Machine Learning Framework for Nonlinear and Interaction Relationships Involving Categorical and Numerical Features

By

Shirish Mohan Rao[1]

Presented to the faculty of the graduate school of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements for the degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF TEXAS AT ARLINGTON

AUGUST 2021

Supervising Committee:

Dr. Victoria C.P. Chen, Supervising Professor

Dr. Atefe Makhmalbaf

Dr. Jay Rosenberger

Dr. Shouyi Wang

[1] Department of Industrial, Manufacturing and Systems Engineering. The University of Texas at Arlington. Arlington, TX 76019 USA Shirish.rao@mavs.uta.edu

Copyright © by Shirish Mohan Rao 2021

All Rights Reserved

Machine Learning Framework for Nonlinear and Interaction Relationships Involving Categorical and Numerical Features

August 31, 2021

Contents

List of Figures

List of Tables

Abstract

All differences in this world are of
degree, and not of kind, because
oneness is the secret of everything.

Swami Vivekananda

Supervising Professor : Dr. Victoria C.P. Chen

Traditionally, physical scientific experiments have been conducted extensively to study and understand the behavior of a process or a system. With the advancement of computing technology in recent years, computer codes and algorithms are used as simulators to replicate behavior of a complex system. Such use of computers to study a system is termed as 'computer experiments.' The process involves selecting specific points or runs in the design space in order to maximize information about the system in minimal runs. These computer models are high dimensional and can take a long time to simulate. Metamodels (or surrogate models) built using the data collected from computer model experiments are hence used to approximate the functional relationship between inputs and outputs.

The contribution of this dissertation falls in design points selection and modeling stages of the above process. First, existing computer experiments with mixed factors (categorical and numerical) are reviewed and then we perform a comprehensive study of these designs to understand their performance under various settings. In the latter part of the thesis, we propose a data-mining framework to learn and model interactions and non-linearity with categorical and numerical features.

ACKNOWLEDGMENTS

Throughout my tenure as a graduate student, I have received tremendous support and assistance. Most importantly, I would like to express my deepest appreciation to my advisor, Dr. Victoria Chen for her tremendous kindness, support, inspiration and patience during my years as a Ph.D. student. I thank her for giving me the opportunity and trusting me with this research. I could not have asked for a better advisor during this program. I would also like to extend my gratitude to my committee members Dr. Atefe Makhmalbaf, Dr. Jay Rosenberger and Dr. Shouyi Wang for their guidance during the research. I would like to recognize the assistance provided by Richard Zercher, Ann Hoang and Cindy Royster from IMSE department at UTA for their technical and administrative support. I am also grateful to my colleagues and friends especially Amith Viswanatha, Nilabh Ohol, Nahal Sakhavand, Alireza Fallahi, Maryam Moghimi, Ashkan Farahani and Priyanka Dadhich for giving me some of the best memories in life. I was extremely fortunate to be surrounded by many 'pawsome' friends during my time at UTA, especially Ariel, Theo, Oliver, Tejo, Arya, Rocky, Zelig, Zuri, Zane and Coco. They provided me with unconditional love and happiness.

Finally, I would like to thank my parents and my brother. None of this would have been possible without their love and encouragement.

1 Introduction

Buildings play a vital role in today's economy and growth of a country. The construction industry is one of the largest consumers of natural resources including water, materials and energy. In a study, buildings and construction account for 40% of global energy use and 36% of energy-related carbon dioxide emissions (US-EIA, 2009). Promoting sustainability in this industry is of utmost importance to reduce the impact on the environment. Achieving sustainability related goals and objectives requires consideration of environmental impact of the entire life cycle of the building i.e. raw material extraction, manufacturing, on-site construction, occupancy, demolition and disposal. Understanding the multifaceted relationship between environmental impacts and building design is the key to achieve sustainability (Wong, 2015) .

An engineer or architect faces multitude of decisions regarding the buildings design in the early stages of the process. Apart from the basic function of the building to provide shelter, buildings have different objectives related to environment, social and economic. These are often conflicting objectives requiring a tradeoff between them. An integrated design process involving the key stakeholders, owners, users, engineers and architects will facilitate in making critical decisions and to successfully design a sustainable building. In recent years, the concept of 'Green Buildings' has been growing, promoting the objective to make buildings more sustainable by efficient use of resources while still maintaining the needs and demands (USGBC, 2018).

1.1 Life Cycle Analysis

Buildings have social, economic and environmental impacts not only during operation, but also during its other stages of life cycle. The different stages of a building life cycle can be categorized as resource or raw material extraction, manufacturing, on-site construction, occupancy, demolition and disposal. The environmental impact and energy usage in every

stage is considered while studying life cycle assessment. Life cycle assessment (LCA) is a qualitative tool used to measure the environmental impact of a product in its entire life cycle. LCA has been widely used in evaluating the impacts over entire life cycle of products in automotive sector, manufacturing industry and consumer product (De Kleine et al., 2011; Keoleian et al., 2004; Spitzley et al., 2005). However, evaluating LCA of a building is challenging and complex because of the buildings long service life. The average service life of a building in the U.S is 120 years (Institute, 2014).

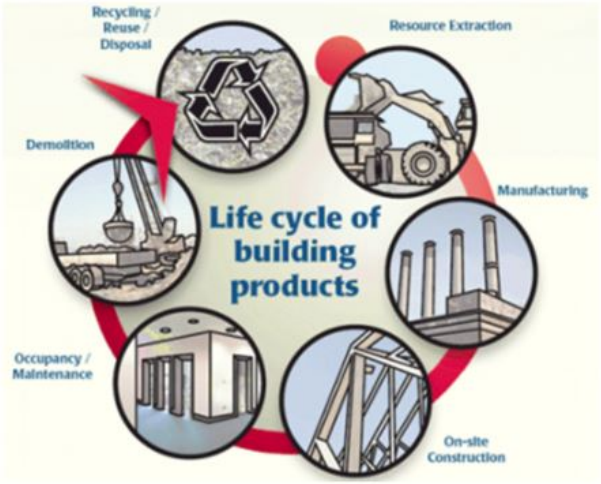


Figure 1.1: Life cycle analysis

The numerous components, assemblies and equipment required for constructing the building from multiple suppliers and contractors adds to the complexity of performing a detailed LCA analysis of a building. Gathering data from these suppliers and evaluating them can be a complex and tedious process. Also, the integration between building simulation software and LCA tools is not very efficient. Most of the building simulation tools focus on the performance of the building in terms of energy consumption, HVAC and lighting design. The common measures used in LCA are global warming potential, greenhouse gas, usage of non-renewable energy source, acidification potential, human health criteria air-mobile and ozone depletion potential. In spite of some challenges, there are benefits of performing LCA for a building. Performing a detailed LCA helps making informed decision at early stages of the building design process so that the building performs as per the requirement. The

performance of building with different design and different scenarios can be compared and an informed decision can be made at early building design stage. The environmental impact of building over its entire life cycle is studied which helps in evaluating any scope for performance improvement during the buildings operational phase. This in turn helps in budget and financial planning for retrofits, renovation and refurbishment. Performing a LCA also contributes to Leadership in Energy and Environmental Design (LEED) certification points.

1.2 Motivation

The engineers and architects have a vast number of building simulation and LCA tools available at their disposal. Some of the common tools used in this industry are EnergyPlus (US-DOE, 2017), eQUEST (US-DOE, 2009) and Athena (ATHENA, 2017). These tools are used to study the performance of building under different design parameters. In recent years, with the growth of machine learning and optimization algorithms, these tools have been utilized to study the input-output relationship for a specific performance objective of the building. Another utilization is to find the optimum settings for building design parameters to minimize (or maximize) a specific objective related to the building. Since these tools are high-dimensional, with large number of inputs, a full combinational run of input settings is not feasible. Hence sampling techniques called design of computer experiments are used to sample points strategically so that maximum information can be obtained in minimal runs.

The tools mentioned often act as a 'black-box.' That means the exact relation between inputs and outputs are unknown. Traditionally, metamodels (or surrogate models) are used to approximate this relationship. It is crucial for the surrogate to model the relationship as accurately as possible in order for it to be used as a prediction function.

The presence of large number of categorical features, along with possible nonlinear relationship and interactions between features, makes the computer design – metamodel process challenging. This research aims to address these challenges. Although this work is motivated by green building analysis study, the framework proposed can be used for broader

applications which involves complex data structures and computer experiments.

The remainder of this dissertation is as follows. Chapter 2 gives a literature of experimental designs and metamodels used traditionally. The limitations and shortcomings of the literature are discussed, thereby leading to the motivation for this research work. Chapter 3 provides a comprehensive study of computer experimental designs capable of handling mix type features (categorical and numerical). Using existing designs, we also propose two new families and their variants of experimental designs. Chapter 4 proposes a machine-learning framework with the ability to model nonlinearity and interactions for complex data structures with categorical and numerical features. We then discuss future work in chapter 5.

2 Literature review

2.1 Computer simulation

With the advancement of computational power in recent years, engineers and scientists use high fidelity computer simulation models to understand the behavior of a system. In many applications like crash simulations, commercial building performance, fluid dynamics and climate change behavior, it is not feasible or practical to perform experiments. This is where computer simulations, being flexible and having the ability to model complex systems is beneficial. Compared to performing actual experiments (where feasible), simulation model offers time and cost savings. Studies conducted using these simulation systems can facilitate making critical decisions during the development phase of the process or product.

Unlike physical experiments, there is no random error or noise in a computer experiment. Apart from being deterministic rather than stochastic, there are many significant differences between physical and a computer experiment. In order to make a physical experiment more robust, the process of randomization i.e. randomly assigning experimental units to treatments is commonly done. If the experimenter is aware of a factor affecting the response, but is not specifically interested in studying the factor, the technique of blocking is applied in a physical experiment. Then there is also replication or running the experiment multiple times under the same scenarios in a physical experiment. The three concepts mentioned – randomization, blocking and replication are methods used to improve the robustness of the experiment making estimates and conclusions stronger. These techniques are not applicable to computer designs because of them being deterministic without any random noise. Running a computer experiment with the same settings will give identical outputs. Hence, the main aim of a computer experiment can be stated as to maximize information by using as minimum runs as possible. This is achieved by using various techniques of space-filling designs and model-based designs (discussed in the next section).

The computer simulation usually requires large number of input parameters (or factors)

to be defined. These input factors interact in a complex manner to calculate the output metric.The nature of relationship between input factors and output measures is unknown, thus making these simulation models a 'black-box' system. A layout of a typical building performance simulation program (like EnergyPlus,eQUEST,etc) is shown in figure 2.1.

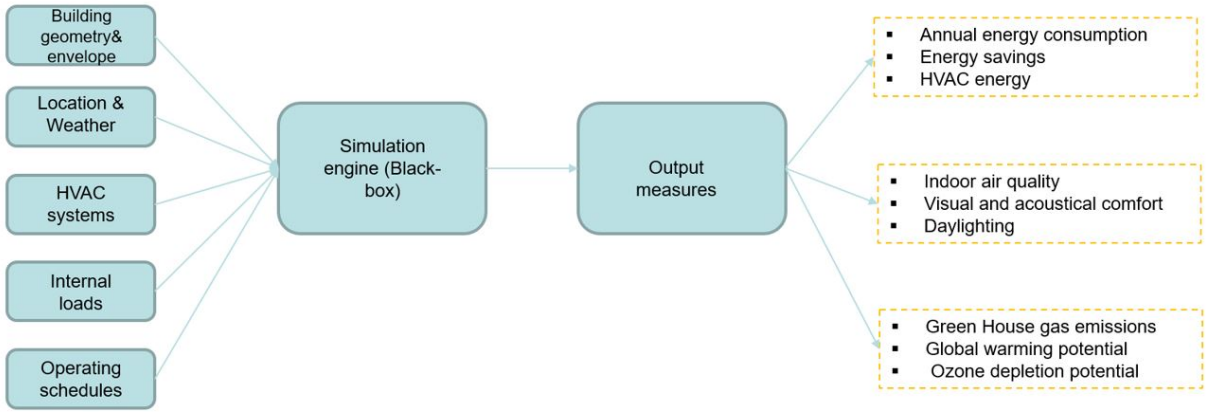


Figure 2.1: Green building simulation example

The input factors can be grouped into high-level categories like location and weather, HVAC systems, building geometry and envelope and internal loads. The output measures like annual energy consumption, indoor air quality and greenhouse gas emissions are calculated by the unknown transfer function using the input factors. In this way, the buildings performance under various scenarios and designs can be studies using a simulation model.

In spite of its numerous benefits, the computer simulation is time and computationally expensive. Since these models are high dimensional, running a single experiment can be computationally intensive. For example, a crash simulation performed at Ford Motor Company took anywhere between 36 to 160 hours to run (Crombecq, 2011). Hence, for efficient utilization of such a simulation model, there are two challenges posed by the methodology of such a study. It is infeasible to run the experiments for every possible setting of the input factor.Hence, careful consideration should be taken in selecting the settings of the input factors to be studied. To optimize this process, experimental design is used. Experimental design gives the settings or 'design points' for the experiments to be run. Ideally, the points

generated from an experimental design should give information of the desired performance metric with minimal runs. The quality of experimental design is crucial to understand the behavior of a complex system. Hence, the first challenge faced by the experimenter is the selection of an appropriate experimental design for the problem at hand. As mentioned earlier, these high dimensional simulation runs can take a long time to run. An iterative, trial and error procedure cannot be used to understand the output measure. To overcome this problem, a metamodel (or surrogate model) is fit. The metamodel is a mathematical function approximating the relation between input factors and outputs. A good metamodel is that which closely mimics the true underlying behavior of the system. The metamodel can then be used to predict the output at unexplored design points (prediction), or to find the settings of input factors to optimize an output measure (optimization). For a metamodel to approximate the output accurately, it is critical the experimental design is selected properly. Apart from time-savings, the metamodel will also aid the experimenter in studying the significant input factors affecting a particular response. This metamodel once built, can be used as a surrogate to the time-consuming computer simulation model. The surrogate is a cheaper and convenient approximation of the actual computer model. Some applications where computer simulation is used are behavorial sciences (Kohli & Peralta, 2017), aerodynamic design (Yondo et al., 2018), turbomachinery design (Peter & Marcelet, 2008), building energy performance (Tian, 2013), and chemical engineering (Diwekar et al., 1992; Diwekar & Kalagnanam, 1996; Diwekar & Rubin, 1994).

We discuss the general process of computer simulation study and then some common experimental designs and metamodels used in the literature.

The general steps in the process of computer simulation study are shown in figure 2.2.

1. Select input factors - The selection of factors to be studied is decided by the experimenter. The input factors selected should be one of interest to the experimenter, and the possibility of the factor significantly affecting the response under consideration. Depending on the type of simulation program used, there might be some factors that

Figure 2.2: Computer experiments - metamodel process

are mandatory for the simulation to run. These types of factors must be included in the design. Since computer simulations can be time consuming, the number of factors to be studied is also a critical part of the experiment. A screening design can be used initially to narrow down the list of few significant factors if the number of factors is large.

2. Experimental Design -The experimental design is a collection of points to run the computer simulation. Ideally, the experimental design should be robust and have good space filling properties so that maximum information regarding the design space can be attained in as few runs as possible.

   An advantage of experimental designs is that the correlations between factors is near orthogonal. This allows the experimenter to make independent estimations of factor effects. Some research have been done regarding this issue. Chapman et al. and Jones (Chapman et al., 1994) suggest to use a sample size of $10d$, where d is the dimension of the computer experiment. Various studies have supported the 10d rule, especially when $d \leq 5$ (Loeppky et al., 2009). However, with high-fidelity experiments, even $10d$ might be too expensive to run. But a more general rule is to evaluate the accuracy of the metamodel, and to add more design points if needed. An advantage of some experimental designs is that the correlations between factors is negligible. This allows the experimenter to make independent estimations of factor effects.

The input factors can be of various types. For example, in the application of building performance simulation, factors like ventilation, ambient temperature take continuous values, and are considered as numerical. The values for insulation and number of windows are discrete numerical, whereas wall type and glass window type are qualitative and categorical. Depending on the types of factors selected, and the number of factors, an appropriate experimental design should be considered. The different types of experimental designs are discussed in the next section.

3. Perform experiments - Using the experimental design, the next step is to run the computer simulations. If multiple processors are available, the simulations can be run in parallel to optimize the time. Depending on the type of computer simulation and its capabilities, the various output measures can be visualized and exported to a spreadsheet.

4. Modeling The goal of an experimental study mainly are: i) to find the optimum setting of factors for the response under consideration,and ii) to understand the relationship between set of factors and the response. A meta-model (or surrogate model) is a model of the high-fidelity computer model, built using the data collected from simulation runs to approximate the response. The metamodel is a computationally quick representation of the computer model. The general relationship between response and input can be written as. $Y = g(x) + \text{error}$,where Y is the response as a function of x. The unknown function $g(x)$ is approximated by the metamodel

$$\hat{g}(x, \beta)$$

where beta is the unknown model coefficients.

2.2 Experimental designs

2.2.1 Classic RSM

Box and Wilson in 1951 (Box & Wilson, 1992) developed response surface methodology (RSM) with the aim to optimize manufacturing process in chemical industry. Responses like high yield, purity and low-cost of chemical reactions were studied and improved using RSM (Dean et al., 1999). A response surface is a geometrical representation of the relationship between factors at different levels and its mean response.

Assuming there are p two level factors, a standard first order design will have 2^p 'factorial points' along with center points. A full factorial $2p$ designs have all the combinations run in the experiment. Hence, with increase in p, the number of experimental runs increases. If a fraction of the full combination is run, then these designs are termed as fractional factorial experiments. A resolution III or higher design is preferred.

A second order design is preferred when modeling using linear effects (first order designs) is insufficient. Estimating quadratic effect is possible using a second order design. The center points and star or 'axial' points augments estimating curvature effect. Central composite designs (Box & Wilson, 1992) and Box-Behnken designs (Box & Behnken, 1960) are some of the most common second-order designs.

2.2.2 Optimal designs

Kristine Smith, a Danish statistician created a family of model-based optimal designs (Smith, 1918). The designs are optimized over a statistical measure based on the model. Depending on the optimality criterion chosen, there are different types of model-based designs. The experimenter has to pre-specify a model for the experiment, and based on an iterative search, these designs are generated by a computer algorithm. In the literature, they have also been referred to as computer-aided designs (NIST, 2009).

Some of the common model-based designs are D-optimality, (which is maximizes the

determinant of information matrix), thereby reducing the variance of parameter estimates), A-optimality (minimizes trace of the inverse of information matrix; minimizes average parameter estimate variance), and G-optimality (minimizes the maximum prediction variance). It is important to know that the design points are dependent on the pre-specified model, these designs are not orthogonal, and hence there may be confounding parameter estimates.

2.2.3 Space filling designs

The computer simulations are deterministic in nature i.e. multiple replications at the same design point will give the same response value. Without a variance component, the model error for a computer design can be attributed to bias. A natural way to reduce the bias is to spread out design points uniformly over the design space. One of the desirable characteristics of an experimental design is for it to give most information with minimum runs. Two design points adjacent to each other will provide similar information. To overcome these challenges, space-filling designs were proposed. The space-filling designs are usually evaluated using a spatial or distance metric.

A sphere packing design is optimized to distance the design points. Johnson, Moore and Ylvisaker defined the maximin and minimax design (Johnson et al., 1990). For a design region D, and with x any point in the design region, the maximin and minimax design can be defined as

$$\max_{\boldsymbol{D}} \min_{i,j} d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \tag{1}$$

and

$$\min_{\boldsymbol{D}} \max_{\boldsymbol{x}\in\mathcal{X}} \min_{i} d\left(\boldsymbol{x}, \boldsymbol{x}_i\right) \tag{2}$$

Along with uniform spread of points, another desirable property of a space-filling design is its projection characteristics on lower dimensions. In practical applications, only a subset of factors have an effect on the response. These factors are called 'active' factors. Hence, when projected on these active factors, the design should have a uniform spread of points.

Latin Hypercube Design, proposed by McKay et al in 1979 overcomes this challenge by placing a constraint while generating the design. LHD requires the number of levels for the factors to be same as the number of runs. Using this requirement, LHD achieves equal spacing of design points in each dimension. LHD is constructed by dividing the design region into evenly spaced cubes, and then sampling only one point across every row and column (similar to Sudoku problem). It is important to know that there are multiple possible LHDs for the same design settings.

Orthogonal arrays, popularized by Genichi Taguchi in the field of quality engineering, is a type of fractional factorial designs. The concept of OA is attributed to Rao (1946) (Rao, 1947) who extended Kishen's (1942) (Kishen, 1942) work of Latin square and mutually orthogonal Latin square. Bose and Bush proposed construction of OA using Hadamard matrices (Bose & Bush, 1952). Other research on the construction of OA are Addelman-Kempthorne (Addelman, 1962) and Rao-Hamming method. An OA of strength t for p variables at q levels, is an arrangement of points in such a way that all level combinations of factors occur in subset of t factors occur with same frequency λ. The general form of OA is

$$\lambda - (n, p, q, t)$$

Use of OAs enables one to study the effect of pairwise combinations of the levels of all the factors in the study. A subset of columns from an OA also forms an OA. This property can be used to sample any number of columns from an already existing OA for the experimental study. The columns of an OA are the factors of study, and the rows represent the experimental run.

An extensive library of existing OAs has been maintained by Neil Sloane (Sloane, 2009). The concept of OA has been extended to incorporate different number of levels for factors. This type of OA has been referred to as asymmetrical OA as well.

2.2.4 Quasi-random sampling

Using discrepancy as a metric, quasi random sampling fills the design points uniformly using a deterministic approach. Discrepancy measures the density of the area in the hypercube, and uses this criteria to space points uniformly. They have been shown to exhibit better space filling property than random designs. Some of the commonly used low discrepancy sequences are Sobol (Sobol', 1967) and Halton (Halton, 1960).

2.3 Metamodel

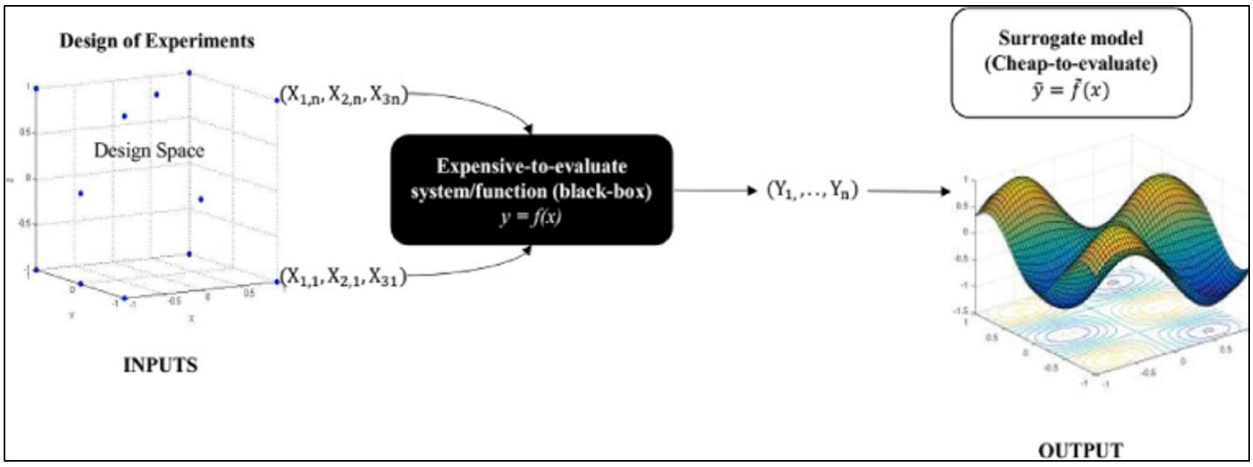


Figure 2.3: Metamodel ((Yondo et al., 2018))

A meta-model is a mathematical representation of the relation between design factors and the response. Many computer models act as a 'black-box' where the characteristics of the underlying function are unknown. A much faster and cheaper meta-model can be built using the data collected from experimental design to understand the relationship and to predict response values of future points. Various types of metamodels have been used in the literature to model system behavior. In this section, we discuss some of the most commonly used metamodels.

2.3.1 Response surface models

The most simple way to fit a surrogate model is to use polynomial regression. Polynomials of any order can be used to approximate the response. The general form can be written as

$$\begin{aligned}\hat{g}(\mathbf{x};\boldsymbol{\beta}) =& \beta_0 + \sum_j \beta_j x_j + \sum_j \sum_{k>j} \beta_{jk} x_j x_k + \sum_j \beta_{ij} x_j^2 \\ &+ \sum_j \sum_{k>j} \sum_{l>k} \beta_{jkl} x_j x_k x_l + \cdots + \sum_j \beta_{j,j,\ldots,j} x_j^d\end{aligned} \tag{3}$$

The model coefficients β can be estimated using least squares since it is a linear model. The size of training data increases as the order of polynomials is increased. Some of the benefits of using a polynomial regression is that they are easy to fit and can be interpreted easily compared to some other meta-models.

2.3.2 Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) was introduced by Friedman in 1991 (Friedman & Roosen, 1995). It is a non-parametric spline based algorithm. The model is built in two stages – forward and backward pass. Model terms in the form of hinge functions are added in the forward pass until a certain stopping criteria is met. In the backward pass, the model terms with the least contribution are pruned. The criteria used to remove model terms in the backward pass is called generalized cross validation (GCV). A MARS approximation can be written as

$$\hat{y}(\mathbf{x},\beta) = \beta_0 + \sum_{m-1}^{M} \beta_m B_m(x) \tag{4}$$

where β_m are the coefficients, and B_m is the basis function. The hinge basis function can be written as,

$$[\pm (x_i - c)]_+$$

where $[\cdot]_+$ is the positive part of the function.

2.3.3 Kriging

Kriging or spatial correlation models is a type of modeling technique, which exploits the correlation between, points to model a response surface. This concept was first introduced by D.G Krige (Krige, 1996). Inspired by this modeling technique, Sacks et al. extended this to be introduced in surrogate modeling (Sacks et al., 1989). The points adjacent to each other will have highly correlated response values, whereas farther points will have less correlated responses. Kriging can generally be stated as

$$\hat{y}(\mathbf{x}, \beta) = \sum_{m-1}^{M} \beta_m B_m(x) + Z(x) \tag{5}$$

With linear model component and a stochastic component. *Z(x)* is a random component with mean zero and covariance. The covariance is given by formula

$$\mathrm{Cov}\left[Z(\boldsymbol{x}), Z\left(\boldsymbol{x}'\right)\right] = \sigma^2 R\left(\boldsymbol{x}, \boldsymbol{x}'\right) \tag{6}$$

where R is called the spatial correlation function. Some of the correlation models are

1. Exponential: $e - \left|\boldsymbol{x}_j - \tilde{\boldsymbol{x}}_j\right| \theta_j$
2. Gaussian: $e^{-\left(\boldsymbol{x}_j - \tilde{\boldsymbol{x}}_j\right)^2 \theta_j}$
3. Linear: $\max\left\{0, 1 - \theta_j \left|\boldsymbol{x}_j - \tilde{\boldsymbol{x}}_j\right|\right\}$

The Gaussian correlation model is the most commonly used in engineering applications.

2.3.4 Radial Basis Function

The general form of radial basis function can be written as

$$\hat{g}(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i \; b\left(\left\|\boldsymbol{x} - \boldsymbol{x}_i\right\|\right) \tag{7}$$

RBF is a linear combination of basis functions and uses the distance (usually Euclidean) between the input space and a point in the design space. Some of the basis functions include linear, cubic, and multi-quadratic. The selection of appropriate basis function is critical for a good approximation. Even though RBF is a linear combination of basis functions, it can approximate nonlinear relations effectively.

2.4 Contribution

The contribution to this research lies in twofold. This research was motivated by defining a framework for optimizing the selection of green building technologies using a computer simulation program. There is abundance of building energy simulation software available today. Some of them are EnergyPlus, eQUEST and Athena. Using a building information model (BIM), which includes detailed information on the building, simulations can be run to understand the performance of building under different scenarios. As mentioned previously, studying the performance of a building using simulations involves different types of factors namely, numerical, discrete-numerical, and categorical. Hence, there is a need to use an experimental design capable of handling this mix of factors. Most of the experimental designs in the literature consider only numerical factors. More recently, experimental designs involving numerical and categorical features have been proposed. However, all of these designs have been evaluated and studied based on their 'space-filling' properties using an appropriate distance metric. None of the experimental designs in the literature are evaluated on the surrogate model built using the designs. In this study, we do a comprehensive study of the experimental designs for mix types of factors. We evaluate the designs based on the surrogate-model's prediction and feature selection capabilities. A surrogate model's performance will only be as good as the quality of data fed to it, and this quality of data will in turn depend on the performance and characteristics of the experimental design used to generate it. Hence, we try to study the experimental designs based on the performance of the meta-model built using it. We also consider a complex mix of non-linear responses

using the experimental designs and meta-model.

The second contribution comes in the modeling aspect of the above process. Certain applications like, sustainability assessment in green building, have a mix of categorical and numerical features. The relation between response and features in these applications can be highly nonlinear in behavior. Moreover, interactions between features impact sustainability metrics, and addressing interaction modeling for this mix of feature types is another challenge. While some of these challenges have been addressed individually in the literature, there is no methodology that handles these complexities simultaneously. We propose a method combining multivariate adaptive regression splines (MARS) with group LASSO to screen relevant features and model terms. Using experimental design, we uncover causal understanding and show that models fitted with our methodology have improved prediction capability.

3 Comparison of Experimental Designs with qualitative and quantitative inputs

Abstract

With the advancement of computational power in recent years, engineers and scientists use high fidelity computer simulation models to understand the behavior of a system. Most of the computer experiments in the literature consider only numerical factors. Many engineering applications also involve a high number of discrete numerical, categorical and ordinal factors. In this work, we propose two different families of experimental designs namely - Kung and Martinez, capable of handling this mix of factor types. We compare their performance with other experimental designs in the literature. The designs are evaluated based on the performance of metamodel (surrogate) fitted using them, rather than a spatial or distance metric traditionally used in the literature.

3.1 Introduction and motivation

Complex physical systems can be studied by using complex computer models. For example, high-fidelity computer simulations closely emulates the real physical system behavior, which enables the experimenter to understand and study the system more efficiently. Many times, performing physical experiments may be too expensive or not feasible at all. A computer model is used in such situations to test the performance of the system under different conditions. Computer simulation models are used in industry to enhance the quality of products and processes and promote innovation. Such simulation models are useful in studying tradeoffs where there are conflicting objectives such as sustainability assessment, which exhibits tradeoffs between, cost and design objectives. Computer models are also widely used in sensitivity analysis and reliability studies. (Tian, 2013)

Most of the literature consider only numerical factors while generating the experimental design. However, the presence of categorical factors in engineering applications is common. Consider the example of a building energy performance simulation. Factors like window type, wall construction type, and building orientation are categorical features with multiple levels. Moreover, there are also discrete-numeric type factors like insulation present in this application. Another example is from solid end milling process where factors like number of flutes, type of work piece material are categorical (Joseph et al., 2020). Such a vast presence of categorical features in engineering applications demands the need of experimental designs capable of handling mixed type of factors. Research in the field of experimental design with categorical factors is very limited. This lack of research can be attributed to various challenges involved in handling categorical factors. Lack of information provided by categorical factors (Ortiz, 2012), their discrete nature and ambiguous center points for categorical factors while using classic experimental design types are just some of the challenges present with the use of categorical factors.

Motivated by the lack of research work in comparing different experimental design types, this paper provides a comprehensive study of performance of these experimental designs.

Although numerous studies have compared experimental designs involving continuous factors using different criteria (Alam et al., 2004; Bursztyn & Steinberg, 2006; Chen et al., 2007), the research on experimental design with mixed factor types is still very scarce. We evaluate the experimental designs on the metamodel fitted using these different design types. Most of the work in the literature evaluate experimental data using a spatial or distance metric. This paper differs existing works because the experimental design is evaluated on the basis of prediction and feature selection. The performance of metamodel will only be as good as the quality of data fed to it, which in turn depends on the type of experimental design. The metamodel is evaluated on prediction and feature selection performance.

The paper is organized as follows. In the next section, we discuss some of the experimental designs found in literature capable of handling mix type of factors. We also discuss two new experimental design types and their different variants. Section 3 describes the simulation study settings, followed by discussion of the results in section 4.

3.2 Experimental designs

3.2.1 Maxpro

Traditional maximin and minimax designs have good space-filling properties in full dimension. However, in most of the experiments, only a few factors impact the response, or we can say that only a subset of factors are 'active.' The traditional maximin and minimax designs when projected onto lower dimension lose their desirable space-filling characteristics. To maximize space-filling properties on projections to all subsets of factors, Joseph et al. proposed maximum projection designs or Maxpro (Joseph et al., 2015). In traditional maximin designs, the points are spread in the feature space so that the minimum distance between them is maximized. Traditionally, the distance metric used is Euclidean. The maximin distance can be represented as

$$\max_{D} \min_{x_i x_j \in D} d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \tag{8}$$

To achieve maximum projection onto subsets of factors, Joseph et al. (Joseph et al., 2015) proposed a new distance criterion. The criterion is defined as.

$$\min_{D} \psi(D) = \left\{ \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{\prod_{l=1}^{p} (x_{il} - x_{jl})^2} \right\}^{1/p} \tag{9}$$

Using a fast derivative based optimization algorithm, the above problem is solved to find the design points.

In 2020, Joseph et al. (Joseph et al., 2015) extended their Maxpro design to accommodate multiple types of factors – namely continuous, nominal, discrete and ordinal. This design is called MaxproQQ (Maximum projection designs with qualitative and quantitative factors) (Joseph et al., 2020). In order to achieve the same, the criterion was changed to

$$\psi(\mathbf{D}) = \left\{ \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{\prod_{l=1}^{p_1} (x_{il} - x_{jl})^2 \prod_{k=1}^{p_2} \left\{ |u_{ik} - u_{jk}| + \frac{1}{m_k} \right\}^2 \prod_{h=1}^{p_3} \left\{ I(v_{ih} \neq v_{jh}) + \frac{1}{L_h} \right\}^2} \right\}^{\frac{1}{p_1+p_2+p_3}} \tag{10}$$

where p_1 is the number of numerical factors, p_2 and p_3 are the number of discrete numeric and nominal factors respectively.

When there are only numerical factors present, the above equation reduces to equation 9. The MaxproQQ design starts with a random Latin hypercube (LHD) design for numerical and discrete-numerical factors. To handle nominal factors, the authors suggests to make use of existing physical experiments literature like fractional factorial designs, orthogonal arrays, D-optimal and I-optimal designs. Using simulated annealing algorithm, the initial design space is then optimized using criterion. The R package "Maxpro" was used to generate

the designs for our study.

3.2.2 Cluster of clusters using Fast Flexible Designs

Most of the space filling designs are generated based on the input space being a hyper rectangle However, many scenarios require the input space to be constrained in a nonrectangular region. Ryan.L and Bradley Jones (Lekivetz & Jones, 2015) proposed a clustering based - space filling design with the ability to create designs for rectangular and non-rectangular region called fast flexible design (FFF). A large sample N is first obtained from the numerical design space. Then, using Ward's minimum-variance criterion (Ward Jr, 1963), n clusters are formed. The experimental design is constructed by using cluster centroid as a design point. Using the above concept of FFF, the work was extended to include nominal factors. This design capable of handling nominal features is called Cluster of clusters (CoC) CoC is initialized similar to FFF design, by a random sample of N in the input space. The design generation can be explained as below (JMP, 2020). Assuming there are m combination of levels of nominal factors, and k design points are allocated to each of these,

1. The N points are clustered into k groups, called the primary clusters.
2. m sub-clusters are formed within the k primary clusters.
3. Within each m sub-cluster, a design point is calculated using the optimality criterion
4. One of the m combination of levels is randomly assigned to each m sub-cluster within each k primary cluster.
5. A design point is chosen using the optimality criteria for each of the m combination of levels, for each k primary clusters. This process is repeated 10 times, or until no improvement is found by changing the given design points.

The optimality criteria used is a weighted version of the Maxpro criterion. The weighted

maxpro criterion is

$$\min_{D} \psi(D)_w = \left\{ \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\prod_{h=1}^{g} I_{w_h}\left(c_{ih} = c_{jh}\right)}{\prod_{l=1}^{p} \left(x_{il} - x_{jl}\right)^2} \right\}^{1/p} \tag{11}$$

where $I_{w_h}\left(c_{ih} = c_{jh}\right) = w_h \geq 1$ if row i and row j have the same value for nominal factor h, and otherwise. For equal weights, the equation reduces to the maxpro criterion of equation 9. For our study, the cluster of clusters design was generated using JMP software (SAS, 2020).

3.2.3 Sliced Latin Hypercube Design

A special type of LHD intended for running computer experiments was proposed by Peter Z.G. Qian (Qian, 2012) called sliced latin hypercube design (SLHD). In this design, a typical LHD is partitioned into smaller slices, where each individual slice is a smaller LHD. This design achieves maximum uniformity in any one-dimensional projection, in any slice. In applications where computer designs are to be run in batches, each slice can be allotted to the different batches of the runs. SLHD can also accommodate nominal factors. An n run SLHD, for p numerical factors can be partitioned into t slices. Assuming t is the number of combination of levels for categorical factors, each combination can be assigned to the individual t slices, thereby forming a complete design accommodating numerical and categorical factors. To improve the computational requirement and efficiency, Shan Ba et al. developed a two-stage algorithm to generate SLHDs. We refer the reader to Optimal SLHD (Ba et al., 2015)for the design generation and optimization process. The R package "SLHD" was used to generate the designs in our study.

3.2.4 Kung

In 2012, Pin Kung studied multivariate, multi-stage green building framework (Kung, 2013). His work utilized building performance simulation software, along with design and analysis using computer experiments (DACE) approach to study building options that impact energy usage and cost metrics. Kung proposed a hybrid Sobol sequence(Sobol', 1967) – mixed orthogonal array (MA) to accommodate the mix of numerical and categorical features. The design points for numerical space is generated using a Sobol sequence. Mixed arrays, which enables to have different number of levels for factors, are used to handle categorical features. These two designs are then combined using a two-factor LHD to form a complete design. A schematic of design generation is shown in figure 3.1. In the above schematic, a 96 point

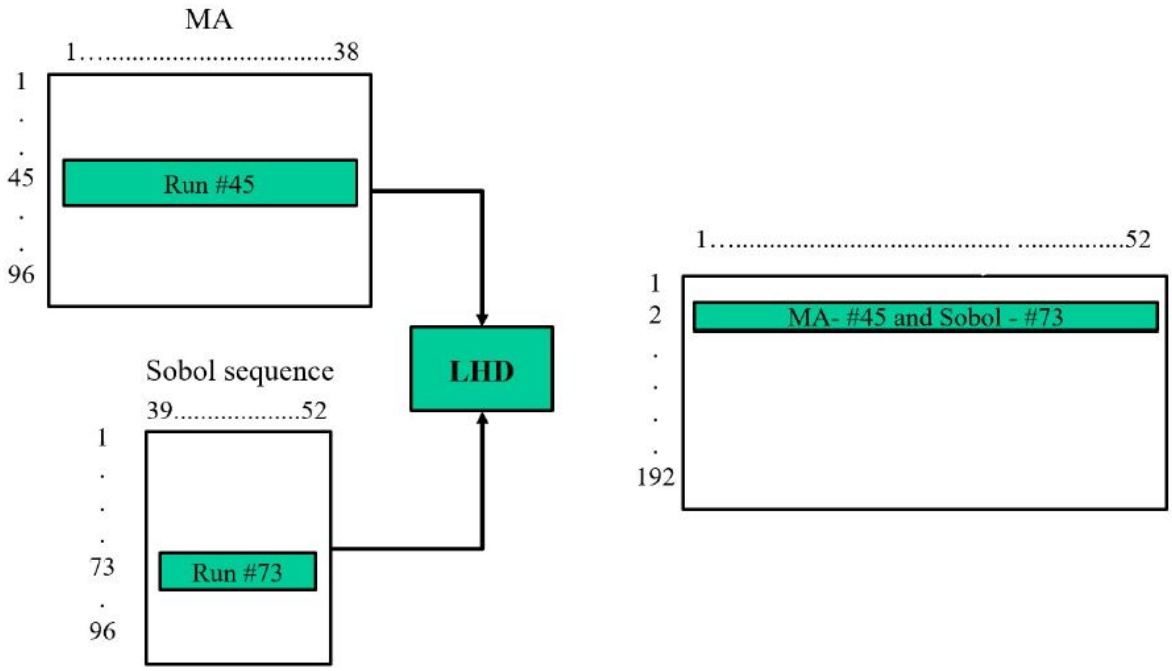


Figure 3.1: Kung's design

MA and Sobol sequence is used for categorical and numerical factors respectively. For the first run, the 45th row of the MA and the 73rd row of the Sobol sequence are combined to form a complete run for the overall design. The numbers 45 and 73 were the first points of a two-factor LHD. The points for numerical space can be generated from any space-filling

design. The steps to construct Kung's design can be stated as:

Step 1: Generate random sample of n points using a space-filling design.

Step 2: Select an appropriate MA/OA of n points.

Step 3: Construct a two-factor LHD.

Step 4: Combine the space filling design and MA/OA to form a complete design.

Based on previously described designs, we create additional two designs using the Kung methodology. The first design is called **Kung- Maxpro or KungMP**, in which we perform step 1 using a Maxpro design for numerical features. The other design **Kung-SLHD** is generated by using a SLHD in the first step of Kung design generation.

3.2.5 Martinez

Nadia Martinez, in her thesis titled 'Global optimization of non-convex piecewise linear regression splines' proposed a categorical adjustment method to handle the mix of variables (Martinez Cepeda, 2013). The design starts in the numerical space within the range [0, 1]. A Sobol sequence is used as the initial design space. The categorical adjustment is done as follows. A 2-level categorical variable takes the first level if the corresponding Sobol sequence value is less than 0.5, otherwise it takes the second level. A threshold is calculated for categorical features higher than 2-levels. The threshold is calculated by

$$\tau = \frac{1}{p - 1\sqrt{p}} \tag{12}$$

where p represents the number of levels. For example, suppose a variable with 4 level has the threshold value 0.6299. If the maximum value of all of the relative values for the variable for each level in Sobol' sequence is equal or greater than 0.6299, the variable takes the level that corresponds to the maximum value, otherwise it takes the last level.

For example, a, b, c and d are the levels of a 4-level factor (X). If $Max\{X_a, X_b, X_c\} = X_b$ then, if

$$X_b \geq 0.62996,$$

then

$$X = b$$

and if

$$X_b < 0.62996$$

then

$$X = d$$

.

Martinez design can be summarized to have the following process:

Step 1: Construct a design for numerical space with predefined columns.

Step 2: Randomize the columns.

Step 3: Perform categorical adjustment to convert to discrete levels.

Similar to Kung's design, we also construct Martinez's design using Maxpro and SLHD in the first step. We call the design **MartinezMP** and **MartinezSLHD** respectively.

3.3 Computational studies

In this section, we consider a comprehensive simulation study to evaluate the experimental designs discussed in the previous section. As compared to the previous experimental designs studies, the aim of our study is to evaluate the performance of designs on the fitted meta-models. The meta-models fitted using these experimental designs will be evaluated on their prediction and feature selection performance.

Along with the previously discussed designs, we also introduce two "benchmark" or designs without "intelligence" for our comparison study. These benchmark designs are generated using the Monte Carlo principle of sampling. For the first benchmark design, which we call Monte Carlo (Uniform), the numerical factors are sampled from a uniform distribution in the region [0, 1]. The categorical factors are then sampled uniformly using integers, which represent the number of levels for that corresponding factor. For example if x_4 is a categorical factor with 4 levels, for a particular run, the level for x_4 is randomly sampled from the set [1,2,3,4]. The numerical and categorical features are then randomly merged to form a complete design. The second benchmark design is similar to the first one, except that the numerical factors are sampled from a beta distribution with parameters α and β set to 1 and 3 respectively. The simulation study has been designed considering various factors discussed below.

3.3.1 Computer model parameters

1. Dimension - It represents the total number of features. The levels for this factor are 12 and 60.

2. Proportion of true features – The screening process is intended to be beneficial in applications with high dimension data where only few of the features are relevant for prediction. Hence, for this study, we vary the number of true features affecting the response to understand the performance of our methodology for different proportions of true features. The levels for this factor are 0.5 and 0.75

3. Response type - We consider three types of interactions in underlying true model. They are

   (a) Type 1 - Numerical-Numerical interaction

   (b) Type 2- Numerical- Categorical interaction

   (c) Type 3- Categorical-Categorical interaction

4. Non-Linear family – One of the motivating reasons to formulate this method was to propose a framework with ability of modeling non-linear relationships. To evaluate our modeling technique on various potential non-linear relations, we refer to the literature of mathematical optimization. In mathematical optimization, there are variety of test or benchmark functions to evaluate the performance of an algorithm. Ideally, these functions should have diverse properties so that the algorithms can be tested for convergence, robustness and general performance (Jamil & Yang, 2013). Derek Bingham and Sonja Surjanovic list a wide range of test functions on their website (Bingham, 2013). From this list of test functions, we summarize and categorize the non-linear relations in three families. The form of these equations and consequently, the levels for this factor are

   (a) NL1 - Non-linear in polynomials up to 4th order

   (b) NL2 - Non-linear in trigonometric functions which includes non-linearity in sin and cos terms

   (c) NL3 - Non-linear in log, exponential and logistic terms.

   The interactions in the true response are allowed to be product of non-linear functions. Example, $f(x_1) * g(x_2)$, where $f(x)$ and $g(x)$ are non-linear. The categorical features in true response are represented using indicator variables. To summarize, the ground truth is of the form $y_i = f(x) + \varepsilon_i$ where $f(x)$ is nonlinear in numerical features space, and also includes categorical and interaction terms.

3.3.2 Experimental designs settings

1. Design Type - The datasets are generated from ten different types of experimental designs capable of handling both – numerical and categorical variables. The design types are namely – Kung, KungMP, KungSLHD, Martinez, MartinezMP, MartinezSLHD, MaxproQQ, CoC, Monte Carlo (Uniform) and Monte Carlo (Beta).

Dimensions	Size	OA	#design points
12	1	L48.2.20.4.9	48
12	2	L80.2.22.4.9	80
60	1	L128.2.100.4.9	128
60	2	L256.2.52.4.3	256

Table 3.1: Selected OAs and size

2. Split - One of the objectives of this simulation study is to understand how the proportion of categorical factors in the designs affects its performance. Hence the two levels for this factor are Split 1 –75% of features are categorical and 25% numerical Split 2 - Equal split (50-50) between numerical and categorical features.

3. Size - The Martinez family (Martinez, MartinezMP and MartinezSLHD) of experimental designs are generated by design for numerical space (Sobol, Maxpro design and SLHD respectively) and then performing a categorical adjustment. On the other hand, the experimental design belonging to Kung family (Kung, KungMP and KungSLHD) and MaxproQQ are orthogonal array based, and hence their size is dictated by the appropriate OA. For this study, we consider two sizes (size 1 and 2) of the experimental design, where size 1 is a smaller design whereas size 2 is a bigger design with more design points. The OA used along with the size for dimensions 12 and 60 are shown in the below table. For an OA with N design points, with $K1$ factors at $s1$ levels and $K2$ factors at $s2$ levels are represented as $LN.s1.K1.s2.K2$

3.3.3 Modeling techniques

We fit the prediction model using gradient boosting machine (GBM)(Friedman, 2001). Apart from having good predictive power, GBM also has the capability of handling mixed type features and ability to model interaction effects. Since GBM is a tree-based model, it performs feature ranking and not feature selection. Hence, for feature selection we use a group regularized LASSO. Yuan and Lin (Yuan & Lin, 2006) proposed the group LASSO penalty that performs grouped feature selection. Group LASSO is useful in settings where categorical

Confusion Matrix		Predicted Model	
		Spurious	True
True Model	Spurious	a	b
	True	c	d

Table 3.2: Confusion Matrix

features are dummy encoded. These encoded features can be grouped together, and group lasso can ensure that all the variables encoding the categorical covariate are included or excluded together (Wikipedia, 2009).

The hyperparameters for boosted tree (number of trees and shrinkage rate) and group LASSO (lambda) are tuned using cross validation. An additional dataset generated by the same type of experimental design as per the case study setting is used as a 'fold' to calculate cross validation error and thereby find the optimum hyperparameter values. The minimum CV criterion is used to perform model selection for group LASSO.

3.3.4 Model evaluation metrics

We evaluate the designs based on their prediction and feature selection performance. For prediction performance, we compare the models using a testing data set generated by MaxproQQ design and calculating Mean Absolute Error (MAE). The formula for MAE is given by

$$\text{MAE} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{13}$$

where $|e_i| = |y_i - x_i|$, y_i is the prediction and x_i is the true value. For feature selection performance of the models, we use a metric suggested by Kubat et al. (Kubat et al., 1998) . A confusion matrix for features can be constructed as shown in table 3.2. This confusion matrix is similar to the type I and II errors of hypothesis testing in statistics.

When there is class imbalance, then accuracy is a highly biased metric and can be misleading. This concept is popularly called as accuracy paradox in the data mining community. Hence, we use a more robust metric suggested by Kubat et.al (Kubat et al., 1998) . From

the confusion matrix, we can calculate sensitivity and specificity as defined by the below formula.

$$\text{Sensitivity} = \frac{d}{(c+d)}$$

$$\text{Specificity} = \frac{a}{(a+b)}$$

Sensitivity or positive accuracy is the proportion of selected true features among all true features. Specificity or negative accuracy is the proportion of unselected spurious variables among all spurious variables (Farahani, 2019). The higher number for sensitivity and specificity is desirable for true model recovery. We calculate the sensitivity and specificity separately for numerical and categorical features.

3.4 Results

The simulation study settings discussed above gave 1440 combinations. Since it is infeasible to show results in plots across all the combinations, we only discuss some interesting results found in our study. Additional plots are provided in the appendix. As mentioned, 100 replications of each case study setting was ran, and based on the selected features from group LASSO, the sensitivity and specificity metric was calculated. A test dataset was generated using MaxproQQ design. We report the testing MAE using the predictions on this dataset. To maintain the uniformity and fairness in comparison, the same columns of selected OA were used for all the OA-based designs (Kung, KungMP, KungSLHD and MaxproQQ).

The horizontal axis of all the plots are ordered in a way that provides easy comparison across the different families of experimental designs. The families can be grouped as Martinez and variants (Martinez, MartinezMP, and MartinezSLHD), Kung and variants (Kung, KungMP, and KungSLHD), designs using Maxpro criterion (MaxproQQ, CoC) and Monte-Carlo based designs (Uniform and Beta).

The boxplot (figure 3.2) shows the prediction across all the cases (split, size, proportion

of true features, response type and non-linearity). The plot shows the performance of all designs in prediction is similar, and there is no statistical difference in the errors. The Monte-Carlo beta design, which is used as a benchmark in this study performed worse than other designs. Yet still not statistically different. This plot demonstrates the challenge in choosing an experimental design for an appropriate application where prediction is of vital importance.

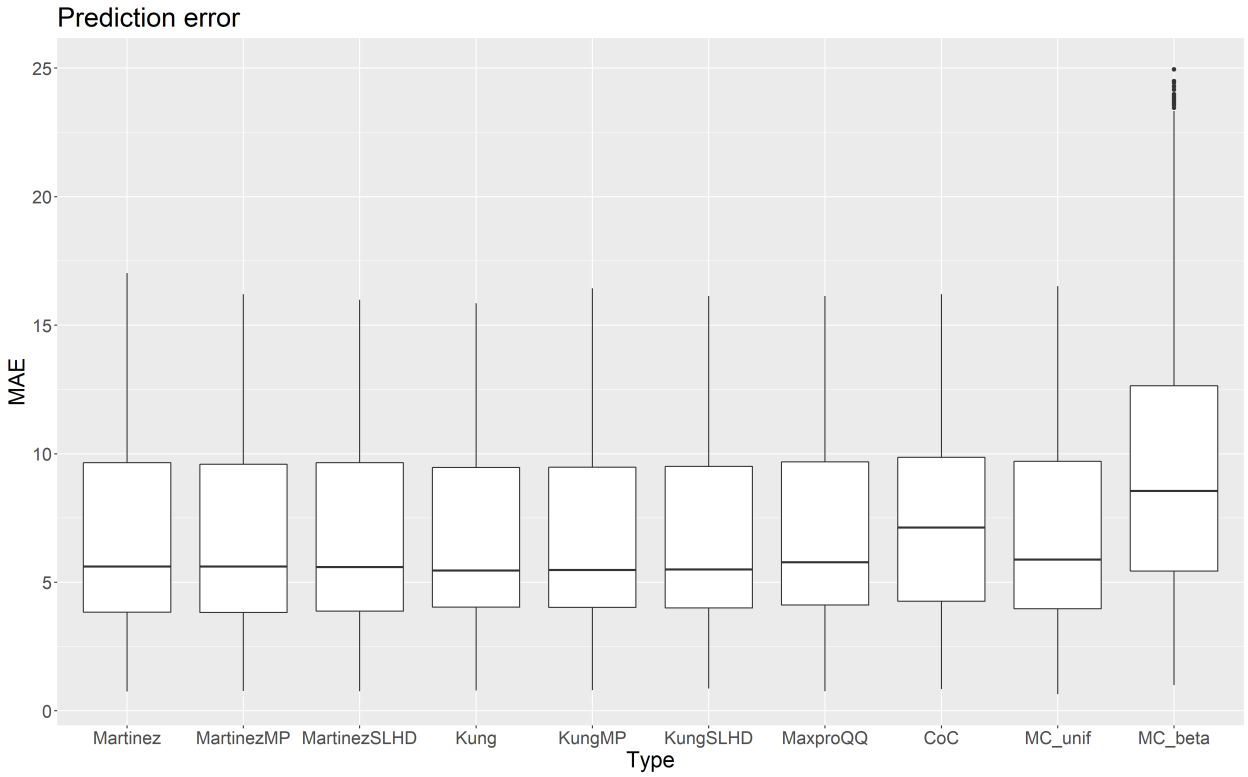


Figure 3.2: Prediction error - dimension 12

The experimental design by JMP demonstrates some superior performance when it comes to identifying the right features, or sensitivity. The figure 3.3a shows sensitivity of numerical factors for responses nonlinear in polynomials. Coc performed highly with sensitivity (numerical) of almost 1, indicating it was able to identify all the factors in the underlying true model. This was followed by categorical adjustment based designs (Martinez and family) and Monte-Carlo uniform. The OA-based designs maintained the sensitivity of slightly less than 0.9. The beta distribution using Monte-Carlo sampling had a poor sensitivity of 0.6. This

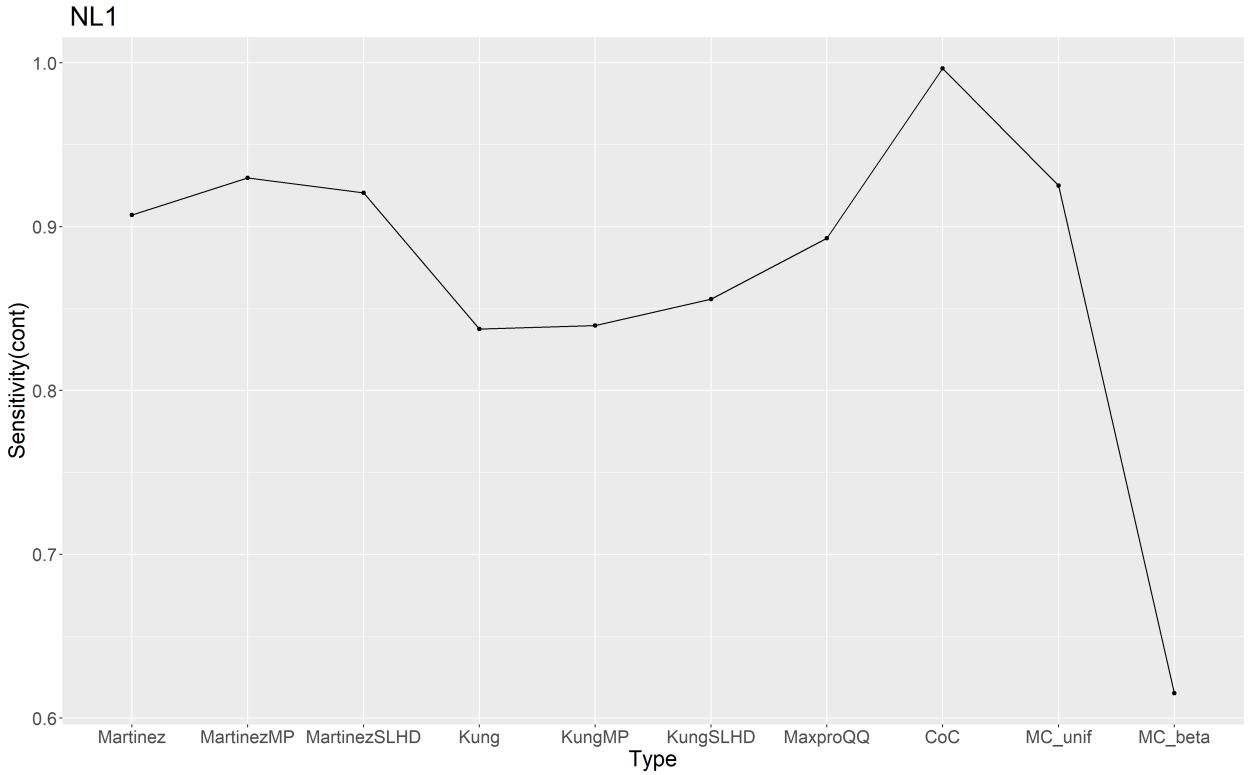
poor performance in selecting the right factors also reflects in its prediction performance. A closer look at sensitivity for this non-linear family is shown in the figure 3.3b. This figure reports the sensitivity only for the response where there is a categorical-categorical interaction (response 3), and the superior performance of CoC is seen. The two lines representing different proportion of true features follow a similar pattern, indicating the experimental designs perform similar to the different proportion of true features.

The specificity for responses 1 and 2 are shown for the second non-linear family (figure 3.4a). The SLHD variants (KungSLHD and MartinezSLHD) demonstrate stronger performance when compared to their other variants, especially when the proportion of true factors is 0.75. The specificity is increased by close to 13% in Martinez design, when SLHD is used to sample numerical space instead of Sobol. This percentage is even higher (about 23%) for Kung design. Another interesting finding is shown in the specificity plot (figure 3.4b), across all the nonlinear family for response type 3. Like sensitivity, CoC also showed higher performance than other designs in specificity of numerical factors.

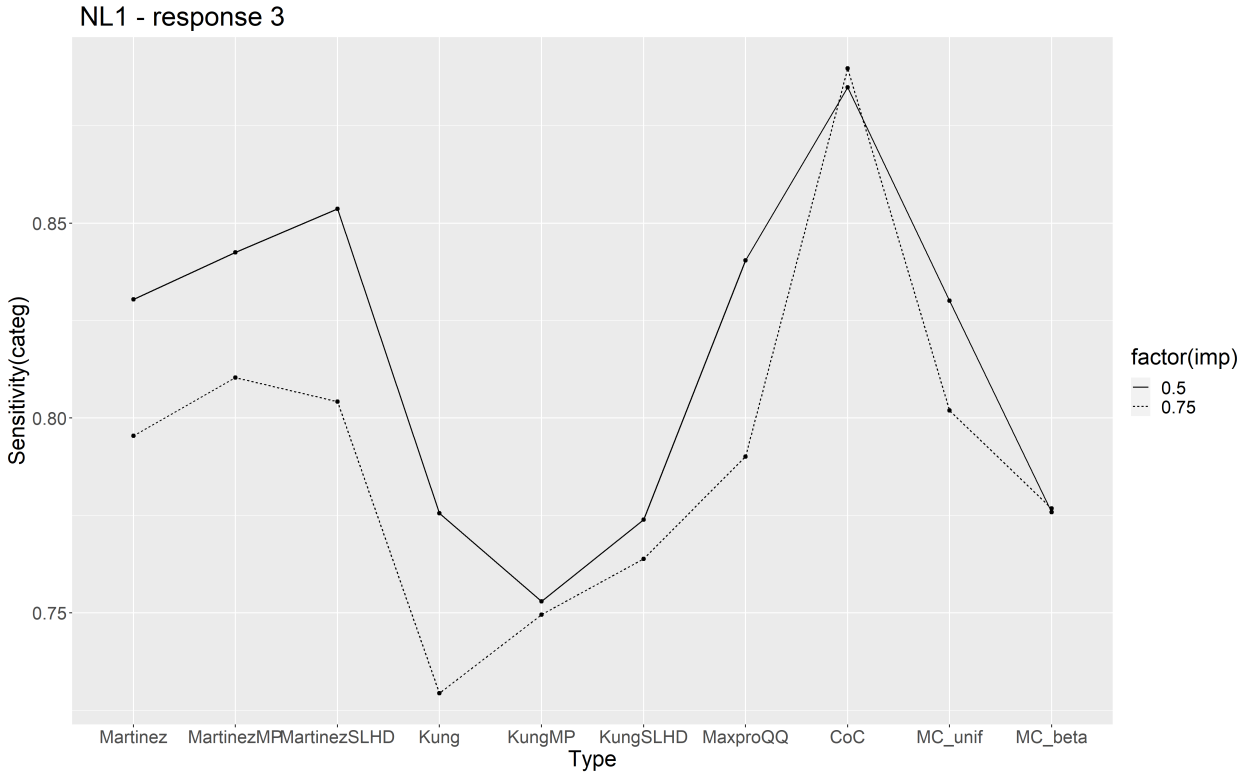
The prediction across all the case study settings are shown in the figure 3.5 Similar to the smaller dimension, the prediction performance of all the designs are identical. As expected, Monte-Carlo (beta) does perform worse.

The OA-based designs exhibit slightly higher sensitivity (categorical) for responses 1 and 3, in the second non-linear family (figure 3.6a). Although the increase in performance is marginal, in applications where identification of the right features is critical, these designs can be appropriate. Under the same response type setting, for non-linearity in exponential and logistic terms (NL3), the Sobol sampled designs (Kung and Martinez) perform better than their variants. (figure 3.6b)

MartinezSLHD shows higher specificity for both type of factors, followed by KungSLHD. Similar to the smaller dimension case study, the SLHD variants have a stronger ability to remove redundant features. In the plot, it is also noticeable that the even though the specificity performance of designs are similar for both the types of factors, the metric is

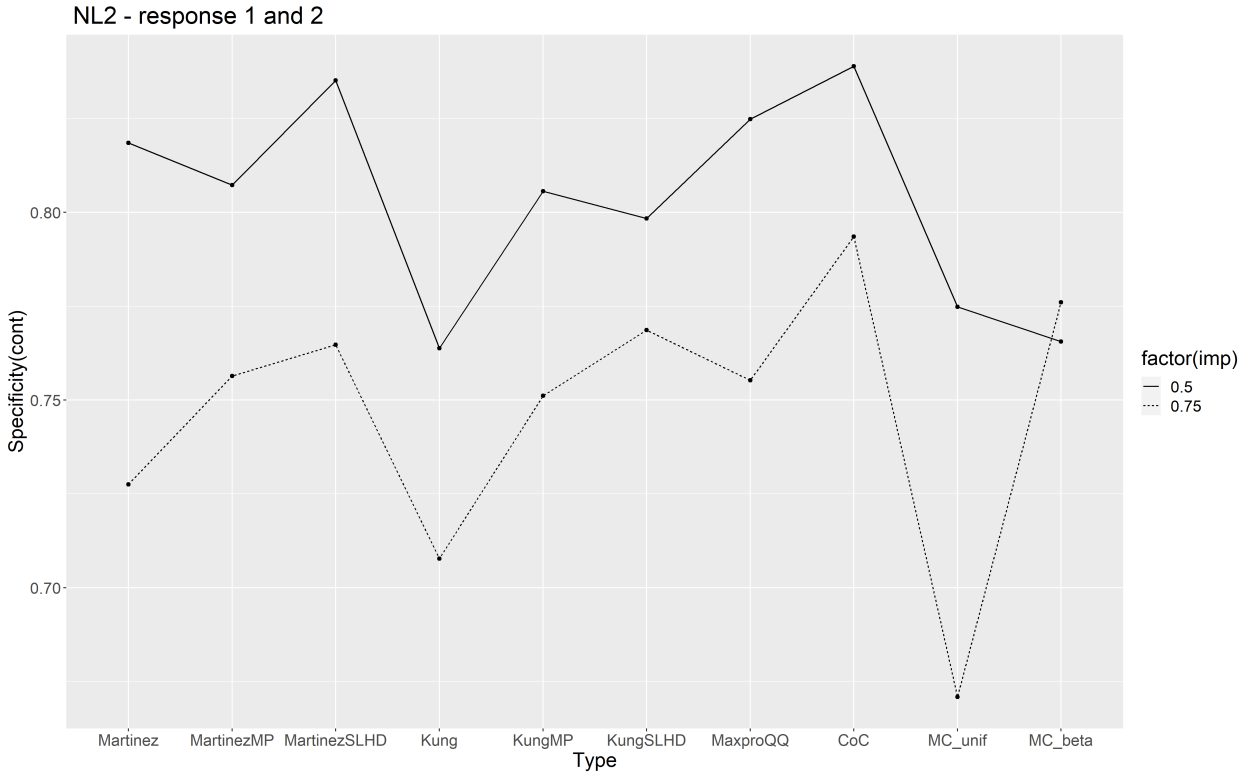


(a) Dimension 12-NL1 sensitivity(cont)

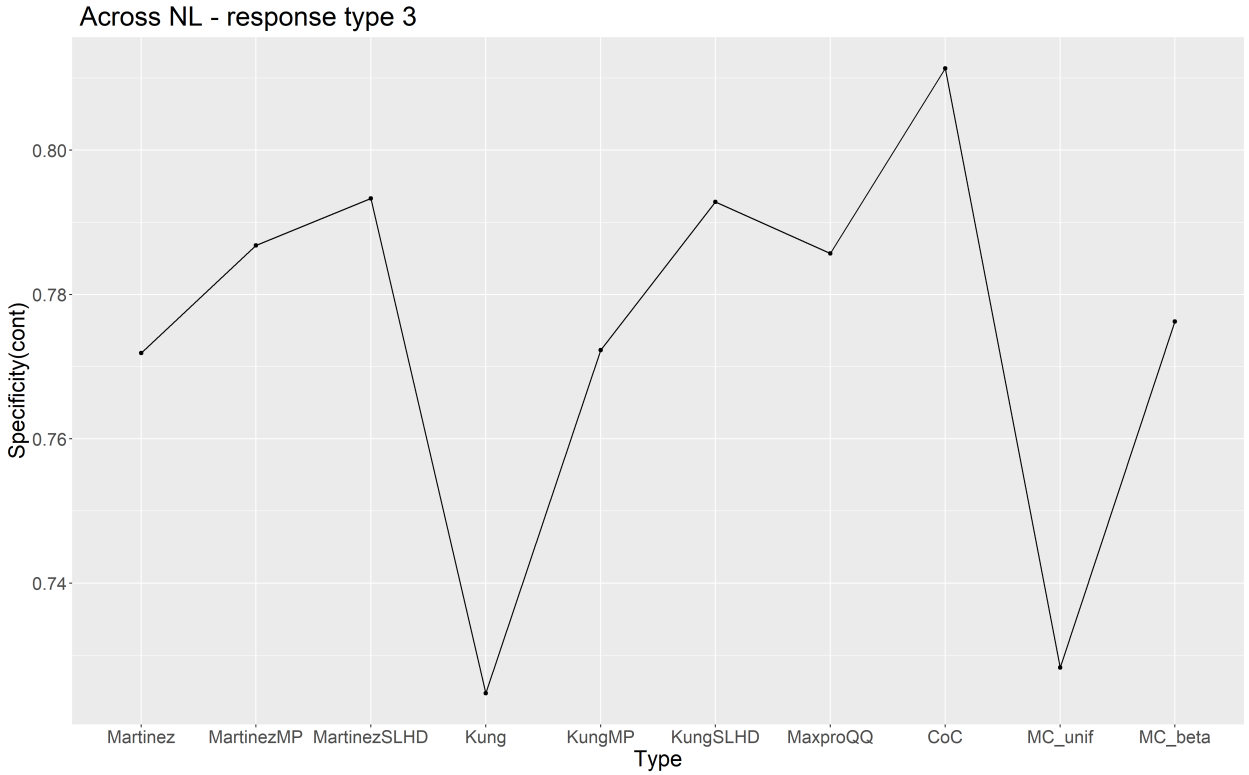


(b) Dimension 12-NL1 sensitivity(cont)-response 3

Figure 3.3: Dimension 12-Sensitivity-NL1



(a) Dimension 12-NL2 specificity(cont)-response 1 and 2



(b) Dimension 12-specificity(cont)-response 3

Figure 3.4: Dimension 12-Specificity

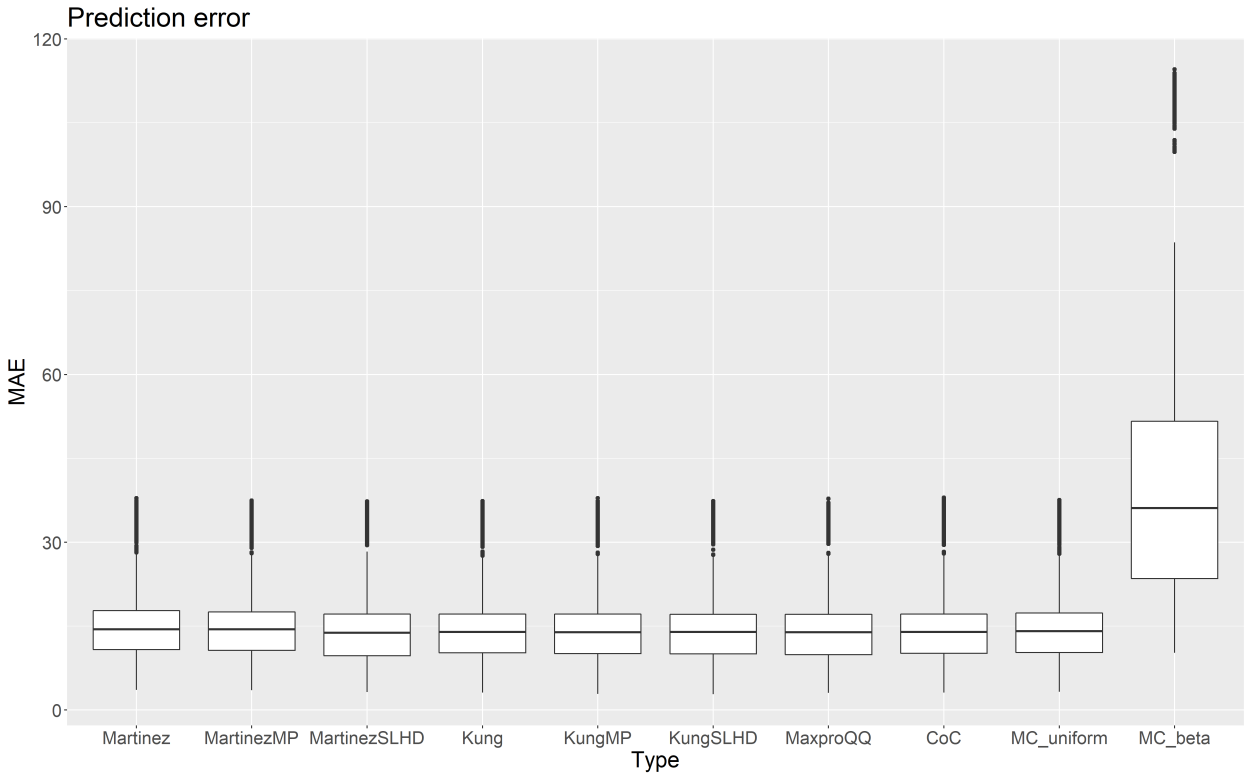
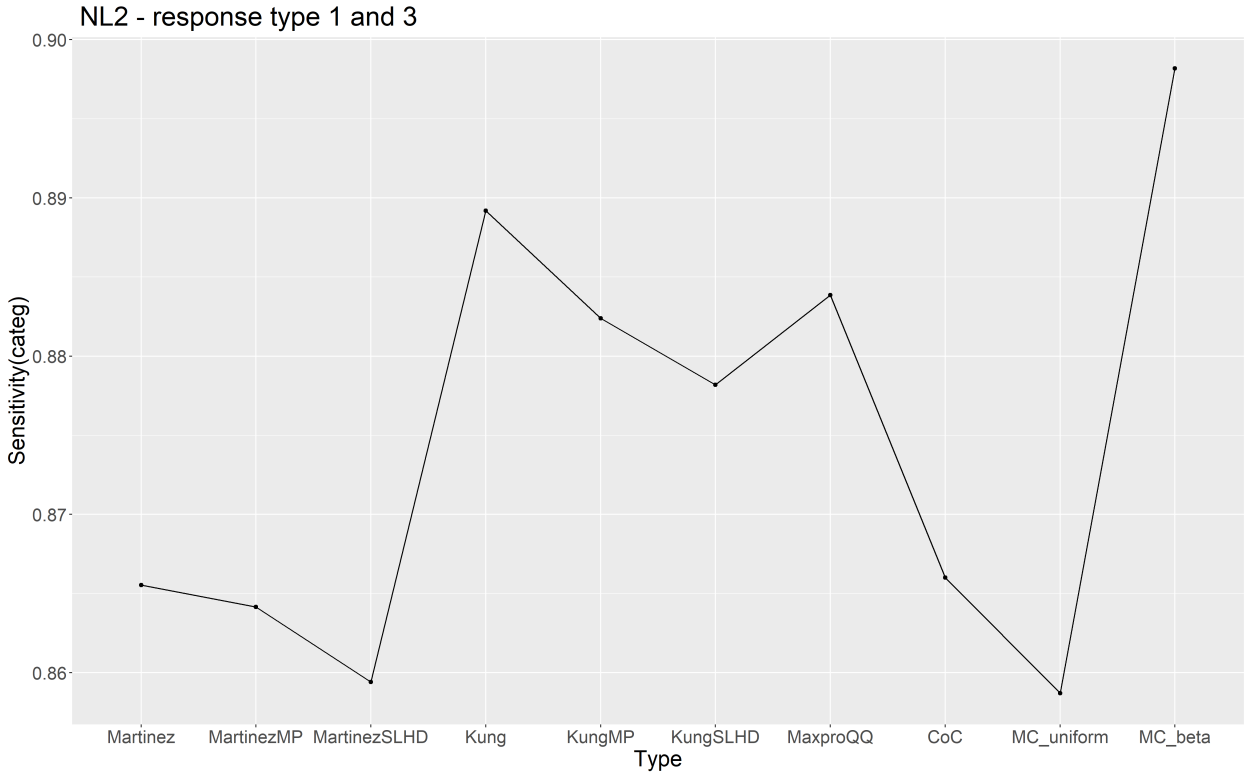


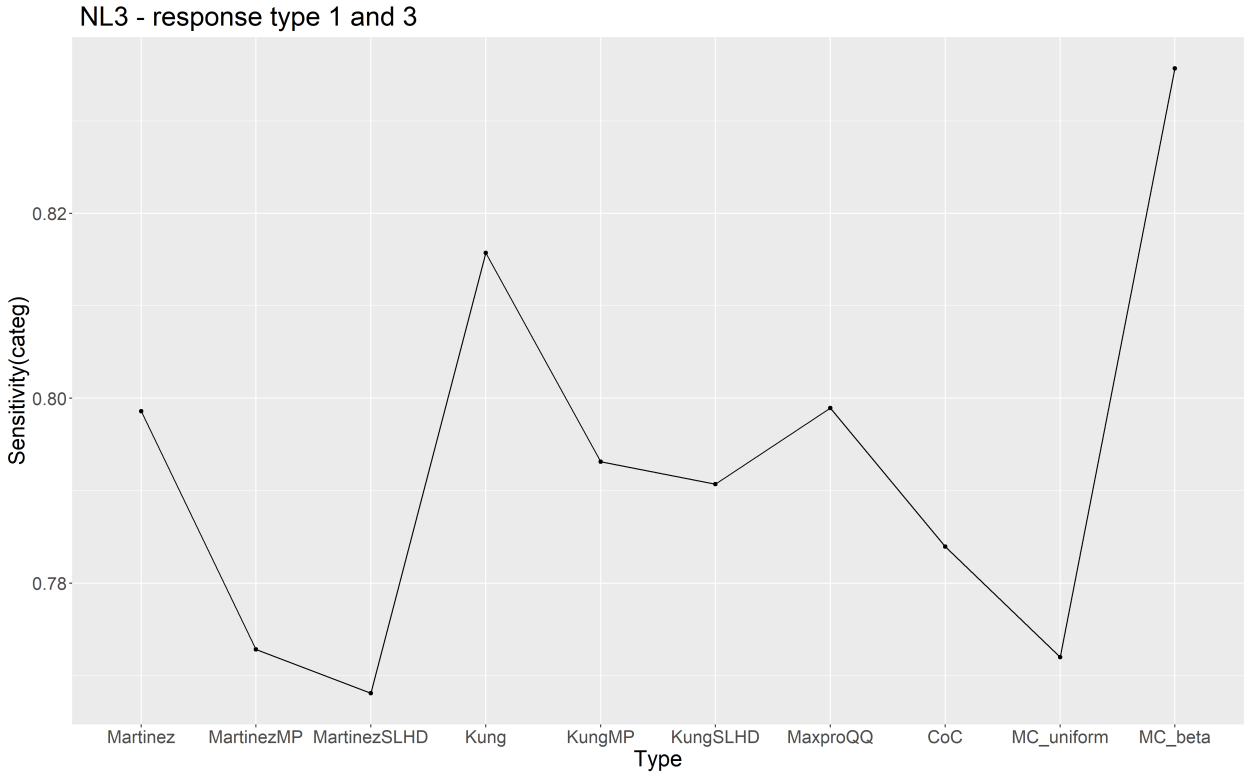
Figure 3.5: Prediction error - dimension 60

scaled down for categorical factors. This indicates the challenges in modeling and feature selection of categorical factors, and that it is easier for group LASSO to perform feature selection on numerical features. (figure 3.7a). The same behavior is also observed for NL3 shown in figure 3.7b.

In surrogate based optimization, the experimental design is the initial step of the optimization process. To optimize a certain process, the algorithm is run in an iterative way, to sample and evaluate the function multiple times. If the generation time of experimental design is slow, the entire optimization process will be inefficient and slow to converge. Hence the time to construct an experimental design is a crucial parameter in an optimization process. The table shows design generation time (in secs) for 12 and 60 dimension designs of size 2. The 12 dimension design has 80 design points and the higher dimension has 256 design points. All the designs were generated using intel(R) Xeon(R) processor.

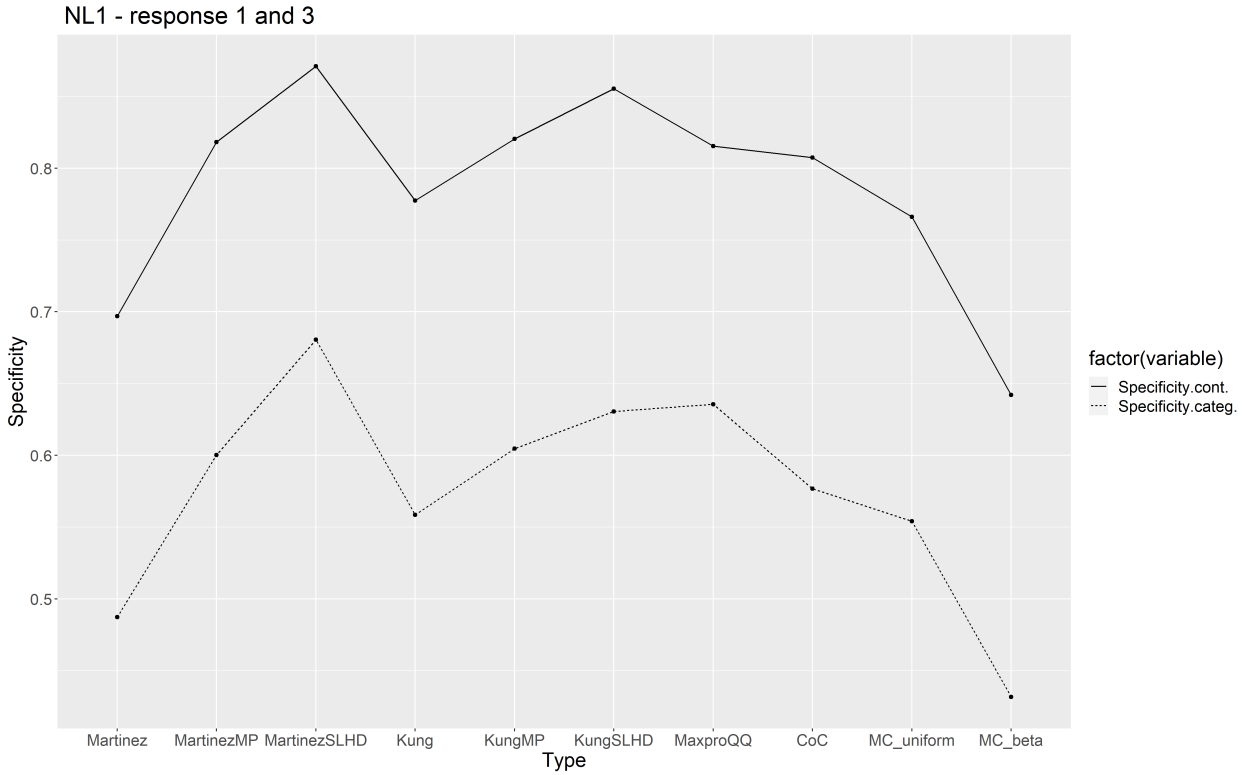


(a) Dimension 60-NL2 sensitivity(categ)-response 1 and 3

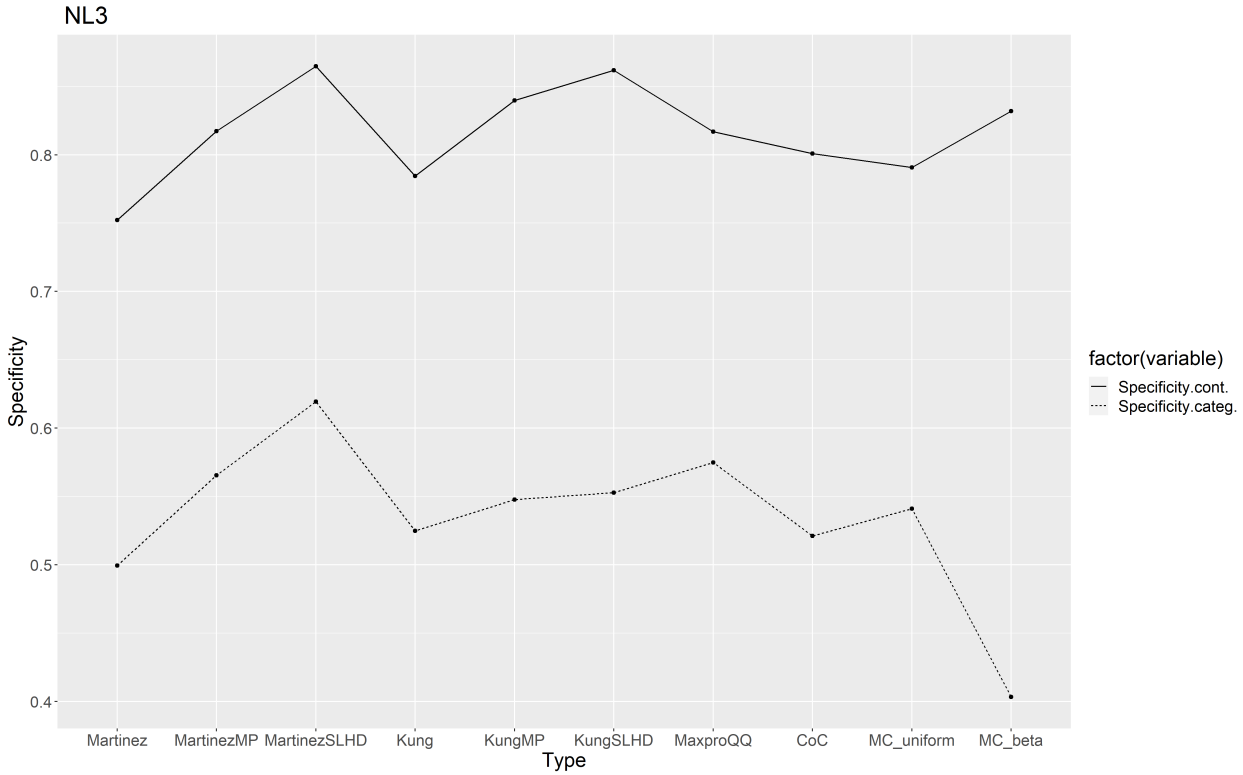


(b) Dimension 60-sensitivity(categ)-response 1 and 3

Figure 3.6: Dimension 60-Sensitivity



(a) Dimension 60-NL1 specificity-response 1 and 3



(b) Dimension 60-NL3-specificity

Figure 3.7: Dimension 60-Specificity

Design type	12 dim	60 dim
Kung	0.205	3.453
KungMP	9.587	375.000
KungSLHD	3.231	102.000
Martinez	0.071	0.007
MartinezMP	18.013	703.560
MartinezSLHD	4.891	285.000
MPQQ	1.897	375.000
CoC	42.350	265.000
Monte-Carlo(uniform)	0.103	0.064
Monte-Carlo(beta)	0.115	0.054

Table 3.3: Design generation time (in secs)

3.5 Conclusion and future work

In this work, we proposed two types of experimental design capable of handling mix type of factors. The proposed designs have the flexibility to change the way they sample numerical factors, thereby creating multiple variants of the design. A comprehensive comparison was conducted on ten different types of experimental designs.The designs were evaluated on the performance of metamodel fitted using the data collected using the designs. While some designs demonstrated marginally better performance than others, there was no design that stood out from other.It would be interesting to study the designs using different meta-models and possibly make some recommendations on design-model combination for a specific application.

4 Machine Learning Framework for Nonlinear and Interaction Relationships Involving Categorical and Numerical Features

Abstract

Certain applications like sustainability assessment in green building have a mix of categorical and numerical features. The relation between response and features in these applications can be highly nonlinear in behavior. Moreover, interactions between features impact sustainability metrics, and addressing interaction modeling for this mix of feature types is another challenge. While some of these challenges have been addressed individually in the literature, there is no methodology, which handles these complexities simultaneously. We propose a method combining multivariate adaptive regression splines with group LASSO to screen relevant features and model terms. Using experimental design, we uncover causal understanding and show that models fitted with our methodology have improved prediction capability.

4.1 Introduction

With developments in data measuring and storing technologies, it has become easier to collect high dimensional data to understand a system.Fields like neuroimaging, bioinformatics, healthcare and finance have taken tremendous strides in using these high-dimensional data for betterment of life around the world in all aspects. This rapid increase of data collection and dimensionality comes with a drawback. Many of the features collected do not significantly affect the system under study, and can be considered as "noise." Sparsity is even more prominent in genomic studies where only a few genes out of millions contribute to a biological outcome (Fan & Lv, 2010). This lead to the development of feature selection concept in machine learning. Sparse models promote less overfitting of the models, and are easier to interpret. The computational power required for a sparser model is significantly less compared to a more complex high dimensional model.The process of finding the subset of features relevant for prediction is called feature selection.

Focusing only on feature selection inducing sparsity, vast amount of literature can be found on feature selection using penalty functions. Some of them are the popular LASSO (Tibshirani, 1996) and its variants – group LASSO (Yuan & Lin, 2006), overlapped group LASSO (Bondell & Reich, 2006; Jacob et al., 2009), adaptive LASSO (Zou, 2006), tree LASSO (Liu & Ye, 2010), smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001), Dantzig slelector (Candes & Tao, 2007), and minimax concave penalty (MCP) (Zhang, 2010) . Group LASSO and mutual information are capable of performing feature selection for categorical features.

In many situations, the main effects are not sufficient to learn the underlying model. This gap can be covered by including first order interactions in the model. When a function $f(x, y)$ cannot be expressed as $g(x)+h(y)$ for some function g and h, then there is an interaction present between x and y (Lim & Hastie, 2015). In a complex system like building performance analysis, the underlying behavior is potentially non-linear. With presence of mixed types of factor, one can assume interactions between features to play a critical part in prediction.

For a dataset with p features, the total number of interactions possible is $\binom{p}{2}$. Hence learning interactions becomes the case of $p > n$, where n is the number of observations. Such a type of problem is suited for penalized regression. Much research has been done to learn interactions using regularization (Bach & Jenatton, 2011; Bach et al., 2012; Lin & Zhang, 2006; Yuan et al., 2009; Zhao et al., 2009).

Categorical features are common in many real world applications. Wall construction, windowpane type and window glass category are some of the examples of categorical features in building performance analysis study. Dummy encoding categorical features will increase the number of features and enforce the "curse of dimensionality." Moreover, interpretation of a model with dummy encoded features might not be straightforward. Of the research discussed above, Glinternet (Lim & Hastie, 2015) is capable of performing feature selection and modeling interaction terms with categorical features. HierNet (Bien et al., 2013) can also model interactions with categorical features, but is limited to only two-level categories.

While above mentioned challenges have been addressed separately, complex data structure like building energy performance analysis demands the need to handle these issues simultaneously. In 2015, David Hsu (Hsu, 2015) used penalized regression to study energy performance in buildings. Different variants of penalized regression – namely lasso, ridge, elasticnet and Glinternet were used as prediction models for the study. Using multifamily and office buildings in New York as a case study, it was found that modeling interactions played an important part in predicting energy performance, and consequently Glinternet performed the best amongst other models. It was also found that out of close to 300 features (mix of categorical and numerical), only 50 of them were actually relevant in predicting energy performance, thereby exhibiting sparsity in features. This work demonstrates the need of a framework capable of modeling interactions as well as screening relevant features, and shows the appropriateness of such a methodology for building performance analysis study.

The next section discusses some of the modeling tools used in the framework. The proposed methodology is then discussed, followed by the simulation study and results.

4.2 Methodology

4.2.1 MARS

Multivariate Adaptive Regression Splines (MARS) is a non-parametric modeling method proposed by Jerome Friedman (Friedman & Roosen, 1995). The modeling is done by combining recursive partitioning on the data and fitting spline functions on each partitioned data. Being a non-parametric modeling method, there is no assumption made about the relationship between predictors and response. Due to this, MARS has the capability of modeling underlying complex non-linear relationships. A MARS model takes the form

$$\hat{y}(\mathbf{x}, \beta) = \beta_0 + \sum_{m-1}^{M} \beta_m B_m(x) \tag{14}$$

where β_m are the coefficients, and B_m is the basis function. The hinge basis function can be written as

$$[\pm (x_i - c)]_+$$

where $[\cdot]_+$is the positive part of the function.

The model building is done in two stages. In the first stage, MARS adds basis functions to the model. The basis functions are selected by the one, which gives maximum reduction in sum-of-squares error. The model terms are then pruned in the backward pass, according to a criterion called generalized cross validation (GCV)

4.2.2 Group LASSO

Regularization is a method used in statistical models to prevent overfitting. To understand the need for regularization in prediction models, the bias-variance tradeoff must be understood clearly. The prediction error of a learning model consists of bias, variance and irreducible error. The relationship between the errors can be seen from the equation below.

$$\text{Err}(x) = \text{ Bias }^2 + \text{ Variance } + \text{ Irreducible Error} \tag{15}$$

The irreducible error is noise in the data, which the current features are unable to explain. The other two errors – bias and variance are controllable and can be influenced. Bias error is the difference between expected value of a prediction model and the true observed value of a particular data point. The bias error depends on the simplicity of the model and the model assumptions. A simple model with fewer assumptions will fail to capture the true underlying relation between the response and the features, also having the tendency to underfit the model, which leads to high bias. Variance error indicates the spread of predicted data from the model of a specific data point. Complex models that fit the training data too well (thus overfitting the model) have lower performance on testing data. These models have higher variance and consequently have higher testing error as well. The 'bias-variance' tradeoff is illustrated in the figure 4.1.

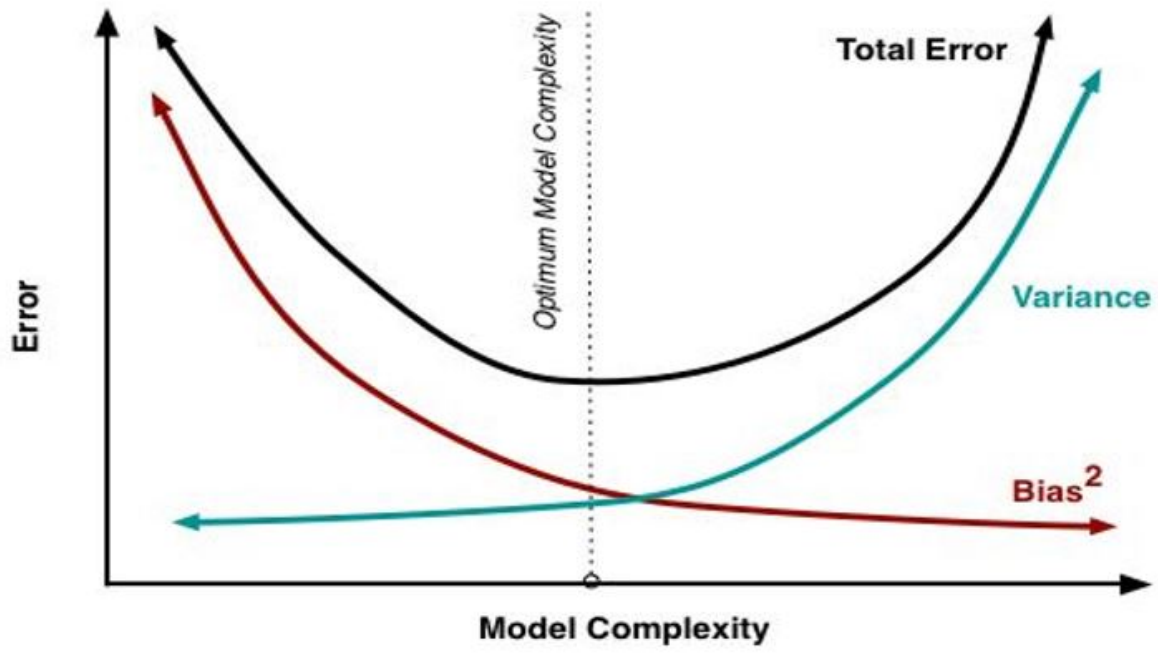


Figure 4.1: Bias-variance tradeoff

Overfitting the model is avoided by using regularization in the model. The regularization introduces a tuning parameter (often represented by lambda) that balances the bias and

variance. The general regression method with regularization can be expressed as

$$|\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \cdot R(\beta) \tag{16}$$

Where R is the penalty or regularization term. The lambda term controls the amount of regularization and shrinks the coefficients. The higher the lambda value, the higher penalty is and coefficients will be driven to zero. Tibshirani (Tibshirani, 1996) proposed Least Absolute Shrinkage and Selection Operator (LASSO), a penalized regression technique that performs shrinkage and feature selection simultaneously. LASSO uses L1 penalty on the regression coefficients, which induces sparsity in the model. The optimization for LASSO can be written as

$$\hat{\beta}^{\text{lasso}} = \operatorname{argmin} \|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{17}$$

In many real- world applications, the features or predictors are grouped or form an intrinsic structure. The underlying structure and relation between the features are known a priori in some applications. The use of structured features is even more prominent in the field of bioinformatics. A challenging problem in this field is to find a mapping between a small subset of loci representing single nucleotide polymorphism (SNP) that impact a family of genes (Bach & Jenatton, 2011; Kim & Statistics, 2012). The genes tend to work in groups and many of these genes share common characteristics thereby having their own structure and have their own underlying hierarchy. Another application in the field of bioinformatics is in diagnosis of tumors. Recent advances in computers and technology, has increased the amount of data available in these fields exponentially. To identify a certain kind of disease, only a small group of genes is relevant out of the thousands of genes present in the human body. Hence, a sparse model improves and helps in interpretability, while reducing the complexity of the model. Sparsity inducing l1 norm and LASSO has been widely used to perform feature selections in these kinds of applications. However, despite the success, LASSO does not consider structured features and the existing relationship between the

features. In neuroimaging and bioinformatics, the features exhibit some underlying intrinsic structure, which LASSO does not make use of in its estimation and feature selection process. The general form of regularized regression can be represented as.

$$\text{Min } f(x) \equiv L(x) + \lambda\Omega(x) \tag{18}$$

Where L(x) is the loss function, is the regularization term and is λ the tuning parameter.

The features form a natural and distinct group in many applications. Yuan and Lin (Yuan & Lin, 2006) proposed the group LASSO penalty that performs grouped feature selection. The groups must be disjoint, which does not allow overlapping in group LASSO. The prior knowledge of groups amongst features is used to improve the performance of the model. Assume the features are in k disjoint groups $G_1 \ldots G_k$, the penalty in group LASSO is

$$\Omega_{\text{gLasso}}\,(x) = \sum_{i=1}^{k} w_i \left\| x_{G_i} \right\| \tag{19}$$

with w_i, the weights for group i.

4.2.3 Boosted Tree

Gradient boosted (Friedman, 2001)tree is a widely used machine-learning technique used for both – regression and classification. Boosting is an ensemble technique, which combines many weak learners in series (typically decision trees), thereby enhancing the prediction performance. Using a loss function (squared loss for regression), the residuals from previous tree are modeled in the succeeding tree, hence improving the model.

The regularization in boosted tree can be controlled by varying the values of model hyper parameters. The critical hyper-parameters associated with boosted tree are the number of trees (weak learners) and learning (or shrinkage) rate. The learning rate is the contribution of each tree by a factor $0 < v < 1$, where v is the learning rate. The two hyper-parameters are dependent, and smaller values of learning rate corresponds to larger number of trees.

It has been found that lower values of learning rate favors better test error. Tree complexity is another hyper-parameter for a boosted tree model. The common practice to find the optimum values of model hyper-parameters is through cross-validation. It reasonable to understand that not all the features have equal importance in a machine learning models. Tree based models are feature ranking rather than feature selection models. Friedman (2001) proposed a formula to measure the relative importance of features in a boosted tree model. The measure takes into account the number of times a feature is selected for splitting, weighted by square improvement of the model, averaged over all trees (Elith et al., 2008). The relative influence is scaled to 100, with higher values indicating more significance of the feature in predicting the response.

We choose boosted tree for our modeling phase for few reasons. Tree based models are capable of handling categorical features, without the need for dummy encoding them. Dummy encoding categorical features increases the dimensionality, making the modeling process complicated. Apart from having good prediction power, boosted tree also has the capability to model interactions naturally.

4.2.4 Glinternet

Using group LASSO as the workhorse, Michael Lim and Trevor Hastie proposed an approach to learn pairwise first order interactions in a regression model. Extending the work of Jacob Bien et al (Bien et al., 2013) titled 'HierNet' , Glinternet (group-LASSO interaction network) models interactions that satisfy strong hierarchy. i.e. an interaction term is added to the model only if the main effects are identified as significant. While HierNet is only capable of handling two-level categorical features, Glinternet can model interactions including categorical features with arbitrary levels. The model for quantitative response is given by

$$Y = \mu + \sum_{i=1}^{p} X_i \theta_i + \sum_{i<j} X_{i:j} \theta_{i:j} + \epsilon \tag{20}$$

with θ_i, the coefficient for main effect X_i, and θ_{ij}, the coefficient for interaction ij. For

computational benefits, overlapped group LASSO (where features can belong to multiple groups) penalty with constraint is solved as an unconstrained group LASSO problem. The constraints are related to the coefficients of different levels for categorical features, where the requirement is the coefficients sum to zero. The methodology was demonstrated on synthetic and real-world datasets, and the results were comparable to existing modeling techniques.

With the modeling tools described, we now describe the proposed screening methodology in the next section. The schematic steps of the proposed process is shown in the figure 4.2

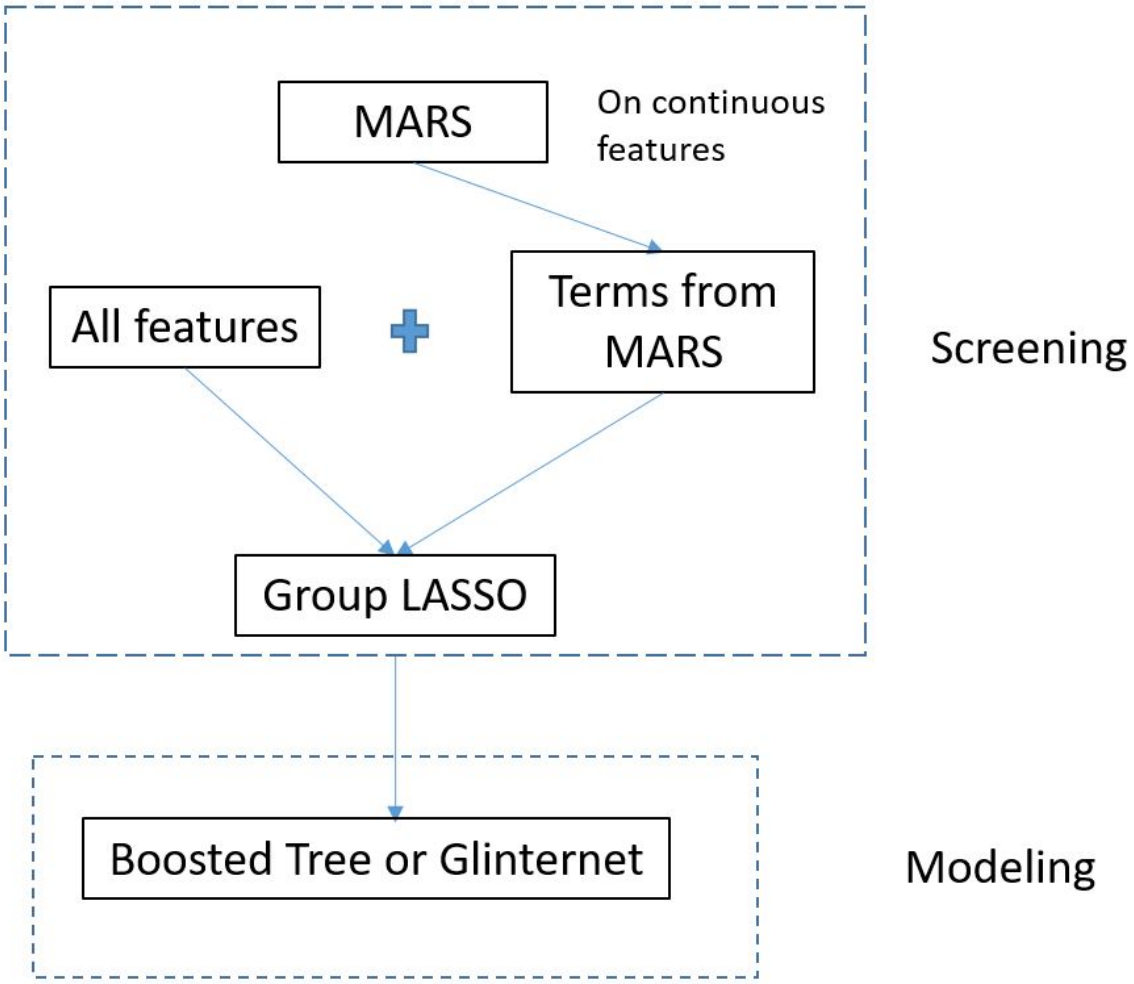


Figure 4.2: Proposed framework

1. Forward Pass using MARS - In the first step, we fit a MARS model only on the

numerical features. The MARS model is allowed to overfit the data by including all the terms in forward pass. The termination criteria for the forward pass related to maximum number of terms is set to be a large number (1000), so that all MARS terms in the form of hinge functions and interaction terms can be included in the model. The backward pass or 'pruning pass' is set to "none" in the MARS algorithm, retaining all the terms from forward pass in the model.

The rationale in using an overfit MARS model in step 1 is to capture non-linear and numerical interactions aspect of the model. The selected terms (hinge functions and interaction terms) from forward pass in MARS are then combined with all features – numerical and categorical. The reason to include all features along with terms from MARS is that in situations where just the linear terms is needed, the model will use these instead of the basis function coming from MARS for prediction.

2. Feature Selection using Group LASSO - The terms from MARS, along with all features are then merged to form the entire feature set. This feature set is then fed to group LASSO to perform feature selection. Group LASSO performs feature selection of groups, by setting non-zero coefficients to important groups of features. The group indices for group LASSO are set in such a way, so that the linear feature along with its corresponding hinge function from MARS (if any) go in the same group. The hinge function is represented as $h(c \pm x)$, where c is the knot position for variable x. Example, a linear feature $x1$ and the hinge function from MARS $h(0.3 - x1)$ get the same group index for group LASSO. For interaction terms from MARS, the terms involving same features are set the same group index. For example, the interaction terms $h(0.58125 - x1) * h(x6 - 0.24375)$ and $h(x1 - 0.58125) * h(x6 - 0.24375)$ gets the same group index due to the same features x_1 and x_6 involved in the interaction terms.

   The group LASSO will thereby perform selection of relevant features as well as inter-

action terms. The hyperparameter selection for group LASSO (λ) is done via cross validation using a dataset generated by experimental design.

The simplest and the most popular way of performing tuning parameter is by performing K-fold cross validation. The data $1, \ldots, n$ is divided randomly into K folds of roughly equal size denoted by $F_1, \ldots, F_k$. For each $k = 1, ..K$, leaving the kth portion out, the LASSO model is fit.

Consider training on (x_i, y_i), $\boldsymbol{i} \notin F_K$,and validating on (x_i, y_i), $\hat{\mathrm{F}}_{\lambda}^{-\mathrm{k}}$ is the estimate on the training set. For each tuning parameter λ, we compute the average error over all folds, (Farahani, 2019). The cross-validation (CV) error is given by,

$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in F_k} \left(y_i - \hat{f}_{\lambda}^{-k}(x_i) \right)^2$

Minimum CV chooses the value of λ that minimizes $\mathrm{CV}(\lambda)$.

Another method for model selection with LASSO is the one standard error rule (1SE). 1SE chooses the most parsimonious model with error no more than one standard error above the minimum CV error (Lu, 2019). 1SE, which results in a sparser and simpler model, is often used in model selection. The lambda value with minimum cross validation error is used for model selection. The reason we use minimum CV as opposed to one standard error rule is that we want to make sure the true underlying model is a subset of features selected by group LASSO with a very high probability.

3. Model fitting - A boosted tree or Glinternet model is fitted using the features and terms selected by group LASSO. Both the models – Boosted Tree and Glinternet naturally model interactions and handle categorical features. To summarize, our screening method uses forward pass in the MARS algorithm to generate potential curvilinear and interaction terms. The backward or 'pruning' pass is done using group LASSO instead of the GCV criteria in the original MARS algorithm, thereby performing feature selection while modeling non-linear and interaction relationships. It is important to note that Any type of prediction model can be used in step 3 instead of boosted trees and

Glinternet.

The proposed framework is beneficial in high dimensional complex data where potential nonlinearity and interactions behaviors are present in the underlying true model. This methodology will screen the relevant model terms and features to fit a prediction model, thereby improving the prediction capability of the model.

4.3 Computational studies

In this section, we consider a comprehensive simulation study demonstrating the methodology discussed in the previous section.

4.3.1 Computer model parameters

1. Dimension - It represents the total number of features. The levels for this factor are 12 and 60.

2. Proportion of true features – The screening process is intended to be beneficial in applications with high dimension data where only few of the features are relevant for prediction. Hence, for this study, we vary the number of true features affecting the response to understand the performance of our methodology for different proportions of true features. The levels for this factor are 0.5 and 0.75

3. Response type - We consider three types of interactions in underlying true model. They are

   (a) Type 1- Numerical-Numerical interaction

   (b) Type 2- Numerical- Categorical interaction

   (c) Type 3- Categorical-Categorical interaction

4. Non-Linear family – One of the motivating reasons to formulate this method was to propose a framework with ability of modeling non-linear relationships. To evaluate our

modeling technique on various potential non-linear relations, we refer to the literature of mathematical optimization. In mathematical optimization, there are variety of test or benchmark functions to evaluate the performance of an algorithm. Ideally, these functions should have diverse properties so that the algorithms can be tested for convergence, robustness and general performance (Jamil & Yang, 2013). Derek Bingham and Sonja Surjanovic list a wide range of test functions on their website (Bingham, 2013). From this list of test functions, we summarize and categorize the non-linear relations in three families. The form of these equations and consequently, the levels for this factor are:

(a) NL1 - Non-linear in polynomials up to 4th order

(b) NL2 - Non-linear in trigonometric functions which includes non-linearity in sin and cos terms

(c) NL3 - Non-linear in log, exponential and logistic terms.

The interactions in the true response are allowed to be product of non-linear functions. Example, $f(x_1) * g(x_2)$, where $f(x)$ and $g(x)$ are non-linear. The categorical features in true response are represented using indicator variables. To summarize, the ground truth is of the form $y_i = f(x) + \varepsilon_i$ where $f(x)$ is nonlinear in numerical features space, and also includes categorical and interaction terms.

4.3.2 Experimental design settings

1. Design Type - The datasets are generated from four different types of experimental designs capable of handling both – numerical and categorical variables. The design types are namely – KungMP, KungSLHD, MartinezMP and MartinezSLHD.

2. Split - One of the objectives of this simulation study is to understand how the proportion of categorical factors in the designs affects its performance. Hence the two levels

Dimensions	Size	OA	#design points
12	1	L48.2.20.4.9	48
12	2	L80.2.22.4.9	80
60	1	L128.2.100.4.9	128
60	2	L256.2.52.4.3	256

Table 4.1: Selected OAs and size

for this factor are Split 1 –75% of features are categorical and 25% numerical Split 2 - Equal split (50-50) between numerical and categorical features.

3. Size - The Martinez family (MartinezMP and MartinezSLHD) of experimental designs are generated by design for numerical space (Maxpro design and SLHD respectively) and then performing a categorical adjustment. On the other hand, the experimental design belonging to Kung family (KungMP and KungSLHD) are orthogonal array based, and hence their size is dictated by the appropriate OA. For this study, we consider two sizes (size 1 and 2) of the experimental design, where size 1 is a smaller design whereas size 2 is a bigger design with more design points. The OA used along with the size for dimensions 12 and 60 are shown in the below table. For an OA with N design points, with $K1$ factors at $s1$ levels and $K2$ factors at $s2$ levels are represented as $LN.s1.K1.s2.K2$

4.3.3 Modeling techniques

To demonstrate the effectiveness of our proposed screening method, we apply screening to Boosted Trees and Glinternet, and compare their performance to models without propsed screening process. Hence the four candidate models in this simulation study are

1. Boosted Trees with screening denoted as 'scrBT'
2. Glinternet with screening denoted as 'scrGLN'
3. Boosted Tree without screening denoted as 'BT'

4. Glinternet without screening denoted as 'GLN'

The hyperparameters for boosted tree (number of trees, shrinkage rate) and for Glinternet (λ) are tuned using cross validation. An additional dataset generated by the same type of experimental design as per the case study setting is used as a 'fold' to calculate cross validation error and thereby find the optimum hyperparameter values. The minimum CV criterion is used to perform model selection for Glinternet. Unlike Glinternet, Boosted tree does feature ranking instead of feature selection.In order to do a fair comparison, we need to variables 'selected' by boosted tree model, rather than variables 'ranked.'

Friedman (Friedman, 2001)proposed relative influence to rank variable importance. The relative influence measure is based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees. Relative influence is scaled between 0 and 100, with higher values indicating larger influence of the variable on the response. For this study, we consider a variable with non-zero relative influence as 'important.'

4.3.4 Model evaluation metrics

We evaluate the designs based on their prediction and feature selection performance. For prediction performance, we compare the models using a testing data set generated by MaxproQQ design and calculating Mean Absolute Error (MAE). The formula for MAE is given by

$$\text{MAE} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{21}$$

where $|e_i| = |y_i - x_i|$, y_i is the prediction and x_i is the true value. For feature selection performance of the models, we use a metric suggested by Kubat et al. (Kubat et al., 1998) . A confusion matrix for features can be constructed as shown in table 4.2. This confusion matrix is similar to the type I and II errors of hypothesis testing in statistics.

When there is class imbalance, then accuracy is a highly biased metric and can be mis-

Confusion Matrix		Predicted Model	
		Spurious	True
True Model	Spurious	a	b
	True	c	d

Table 4.2: Confusion Matrix

leading. This concept is popularly called as accuracy paradox in the data mining community. Hence, we use a more robust metric suggested by Kubat et al (Kubat et al., 1998) . From the confusion matrix, we can calculate sensitivity and specificity as defined by the below formula.

$$\text{Sensitivity} = \frac{d}{(c+d)}$$

$$\text{Specificity} = \frac{a}{(a+b)}$$

Sensitivity or positive accuracy is the proportion of selected true features among all true features. Specificity or negative accuracy is the proportion of unselected spurious variables among all spurious variables (Farahani, 2019). The higher number for sensitivity and specificity is desirable for true model recovery. We calculate the sensitivity and specificity separately for numerical and categorical features.

Along with sensitivity and specificity, we also calculate false discovery rate (FDR). FDR is the proportion of false features among all the predicted features. FDR was used by (Lim & Hastie, 2015) to evaluate performance of Glinternet model. It is calculated by the formula

$$\text{FDR} = \frac{b}{(b+d)}$$

A lower value for FDR indicates smaller proportion of false features selected.

4.4 Results

Every case was simulated 100 times, and the corresponding MAE, sensitivity, specificity and FDR were calculated. The test dataset is generated using Martinez design and the testing MAE is calculated based on the prediction on this dataset. We calculate the sensitivity and specificity separately for numerical and categorical variables to understand the performance of the model on these different types of features. The feature selection metrics are visualized using line plots, and the numbers on the plots are the average over 100 replications. The prediction error is visualized using box plots to see the distribution of prediction errors.

The prediction error across the parameters - split, size, response type and non-linearity are shown in figures 4.3. The different grey shaded boxplots represent the levels for the factor 'proportion of true factors', wherein the darker shade represents 0.25 and the lighter ones indicate 0.5 and 0.75 proportion of true features. The metric used for comparing prediction error – MAE is a scale-dependent accuracy measure. The model with different proportion of true features will have responses in different scales. Hence, a fair comparison would be comparing boxes of the same grey shade for different models. The different rows represent the four types of experimental design under consideration – namely KungMP, KungSLHD, MartinezMP and, MartinezSLHD.

A quick glance at the three plots demonstrates the improved prediction performance of our screening process as compared to the same model without screening. The screening process seems to enhance the performance of Glinternet more than boosted trees. It is noticeable that screening with Glinternet performs better than other models uniformly in all the scenarios.

A comparison of MAE values across the rows for different 'design types' shows that there is no statistical difference between the experimental design types considered for this study. This is consistent with the results in chapter 3, in which there were no statistical difference in prediction for the different experimental designs under consideration.

The sensitivity for numerical and categorical features are shown in figure 4.4. In our

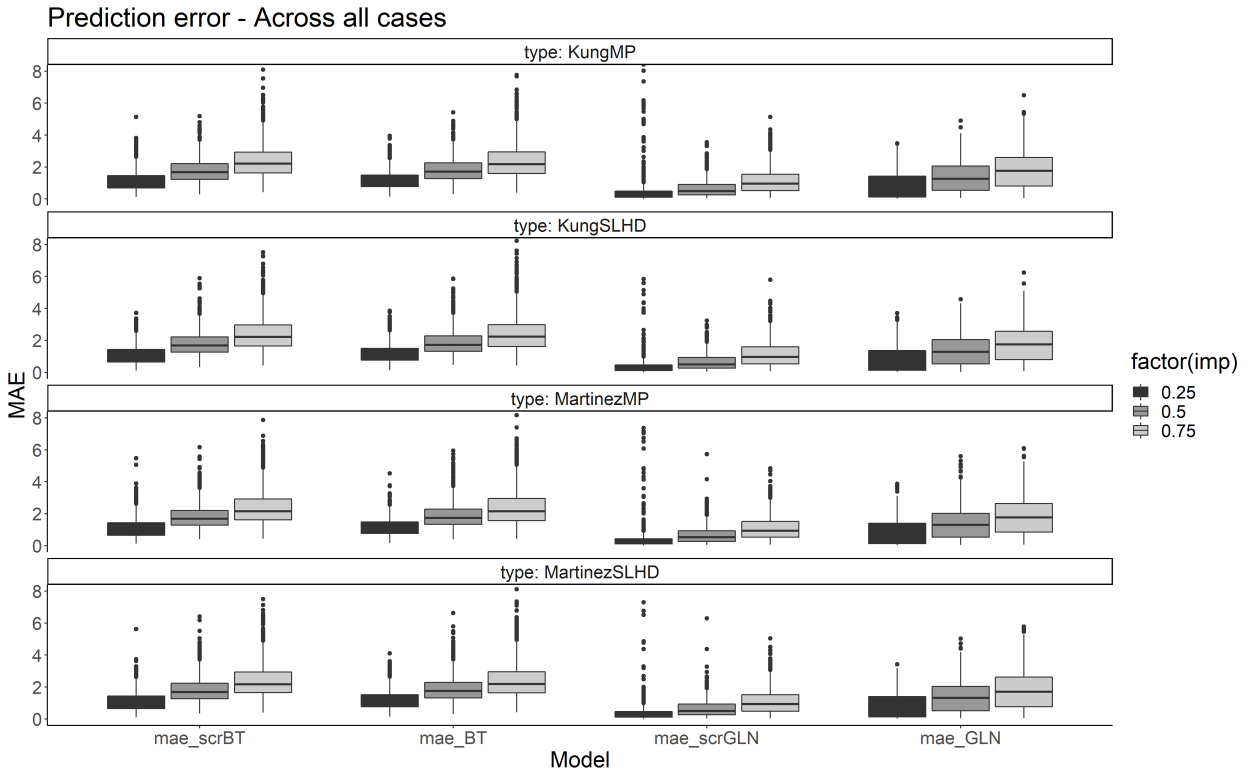
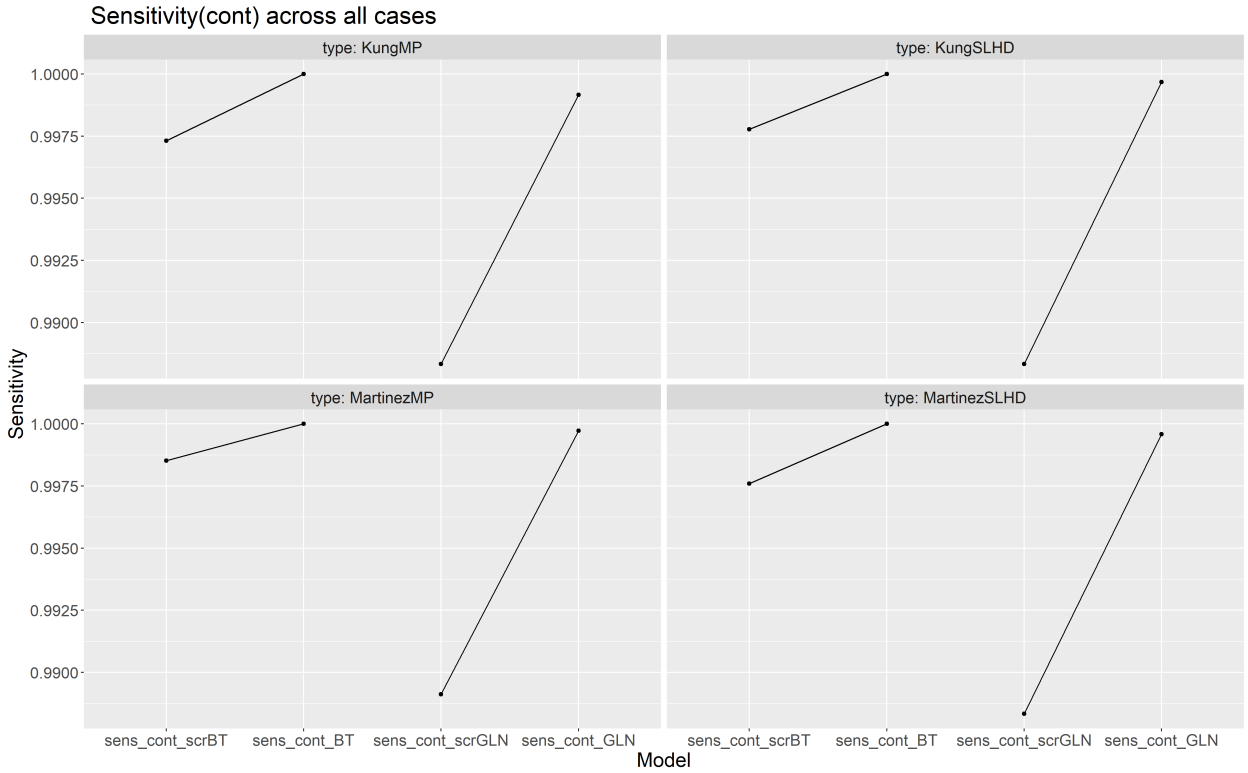


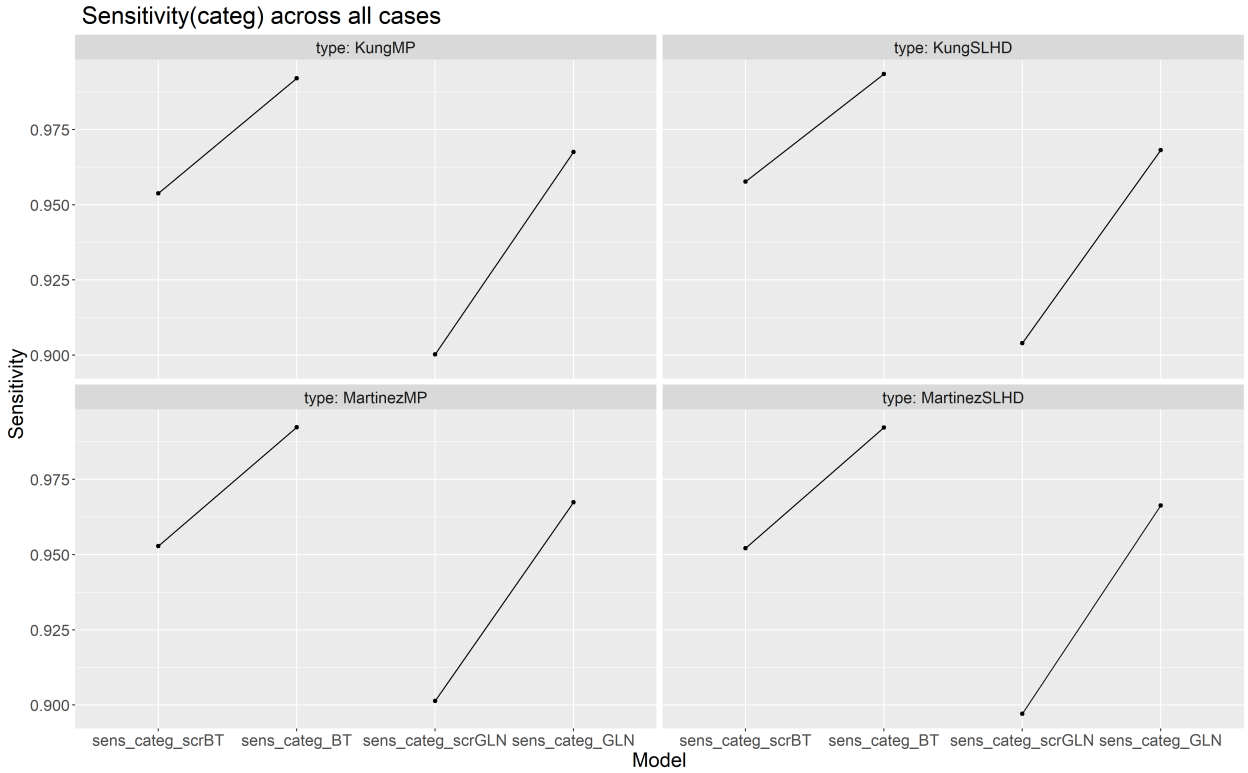
Figure 4.3: Dimension 12 - prediction error

simulation studies, the models without screening tend to select a higher number of features as important, thereby increasing the sensitivity. We can also note that on average, the sensitivity is higher for numerical features. Similar to the prediction performance, there seems to be negligible effect of the four experimental designs on sensitivity.

There is a trade-off between sensitivity and specificity metric of feature selection. A higher sensitivity indicates that there is a potential for the model to select redundant features, along with relevant ones. As mentioned, the models without screening tend to pick a large number of features as important, thereby increasing the sensitivity, but at the cost of specificity. This is illustrated in the figure 4.5a and 4.5b. The proposed framework performs feature selection of main effects and interaction terms using group LASSO. This 'weeding out' of unwanted features is reflected in higher specificity demonstrated by the models fitted with the screening process. The specificity is across the simulation study parameters of split, size, proportion of true features, response type and nonlinearity. We can also note that the



(a) Dimension 12-sensitivity(cont)



(b) Dimension 12-sensitivity(categ)

Figure 4.4: Dimension 12-Sensitivity

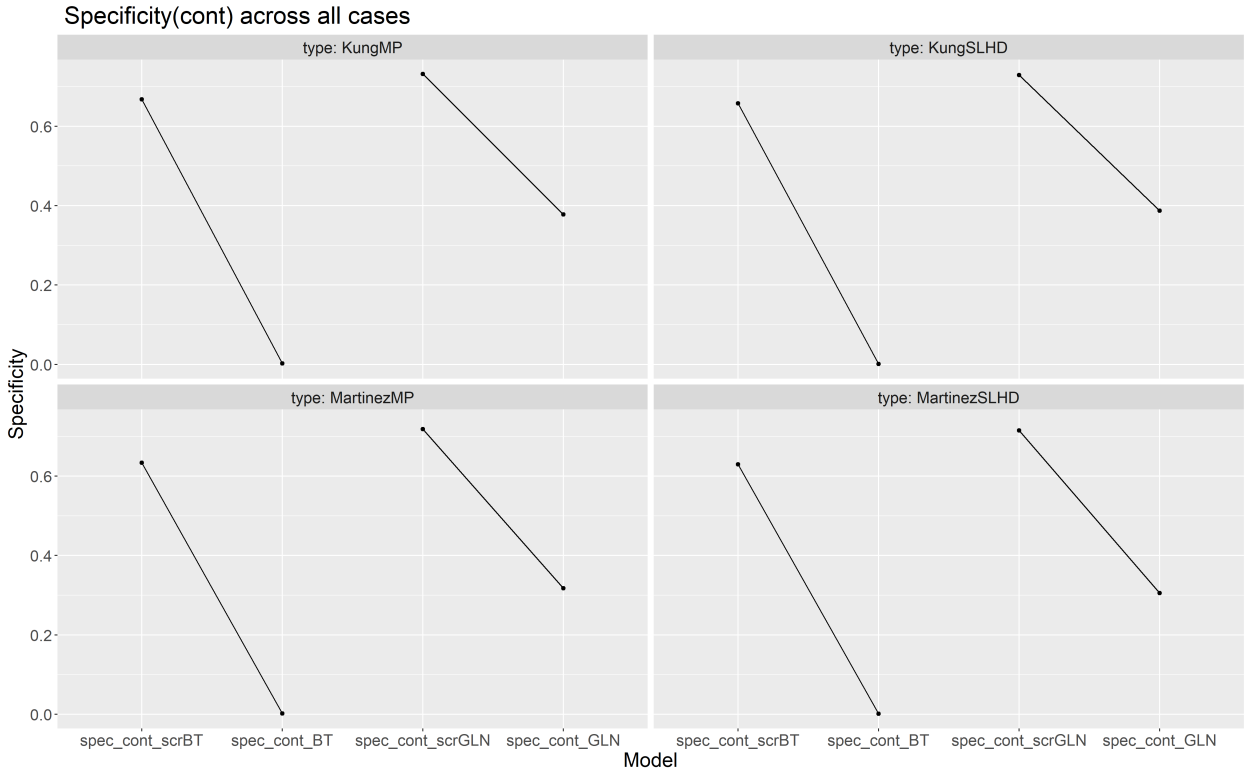
boosted tree models had a lower specificity for categorical than numerical features. The false discovery rate, which is closely related to specificity metric, is shown in figure 4.5c. A lower value of FDR indicates a smaller proportion of false features selected by the model, and the models fitted with our proposed method shows a lower proportion of false features selected. (figure 4.6).

One of the simulation study parameters are the types of experimental designs. The motivation behind including this parameter was to study the effect of experimental designs on the prediction model. It is quite evident from the above plots that this effect is minimal or negligible. To illustrate that, we plot the same test performance values of the models, but with experimental designs on the horizontal scale. The plot shows the interaction between experimental design and model performance. From the figure 4.7, it is seen that experimental design have no effect on either the MAE or FDR of the models. The parallel lines indicate there is no experimental design effect on the models.

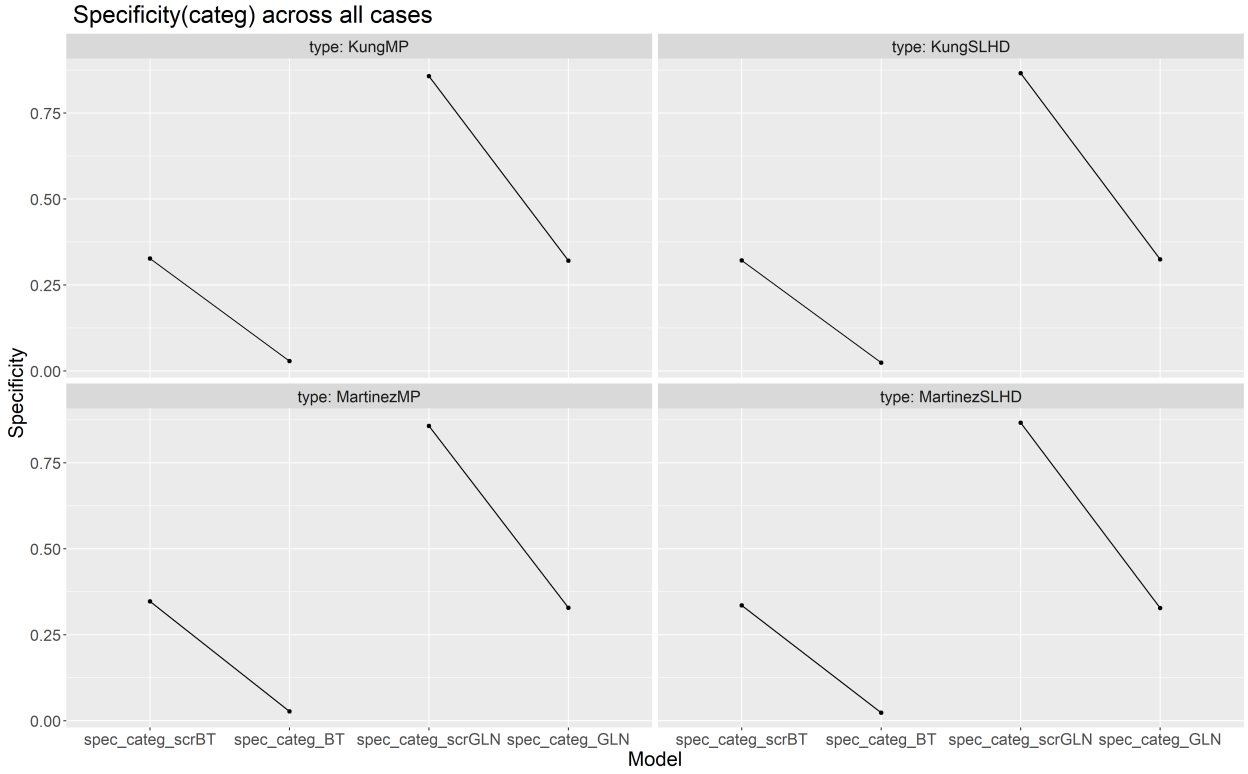
Experimental designs can take a long time to generate depending on the algorithm behind their construction. This time increases with the dimension or number of features. To make efficient use of computing power, we generate only MartinezSLHD type design for the 60-dimension case study.

As in 12-dimension study, the superiority of models with proposed screening method is observed in figure 6. The prediction errors are crossed across all the simulation study parameters, except the proportion of true features.

Similar to the small dimension case study, the screening models show higher specificity, at the cost of slightly reduced performance in sensitivity for both numerical and categorical features. These plots are shown in figures 4.9, 4.10 and 4.11. We can also see that the sensitivity for case with 0.75 true proportion is higher than other levels, because the pool of true features is bigger, making it easier for the models to identify true features. On the other hand, the specificity is higher when only a quarter of features are in the truth, because of large number of features being irrelevant, making it easier for the models to remove redundant



(a) Dimension 12-specificity(cont)



(b) Dimension 12-specificity(categ)

(c) Dimension 12-Specificity

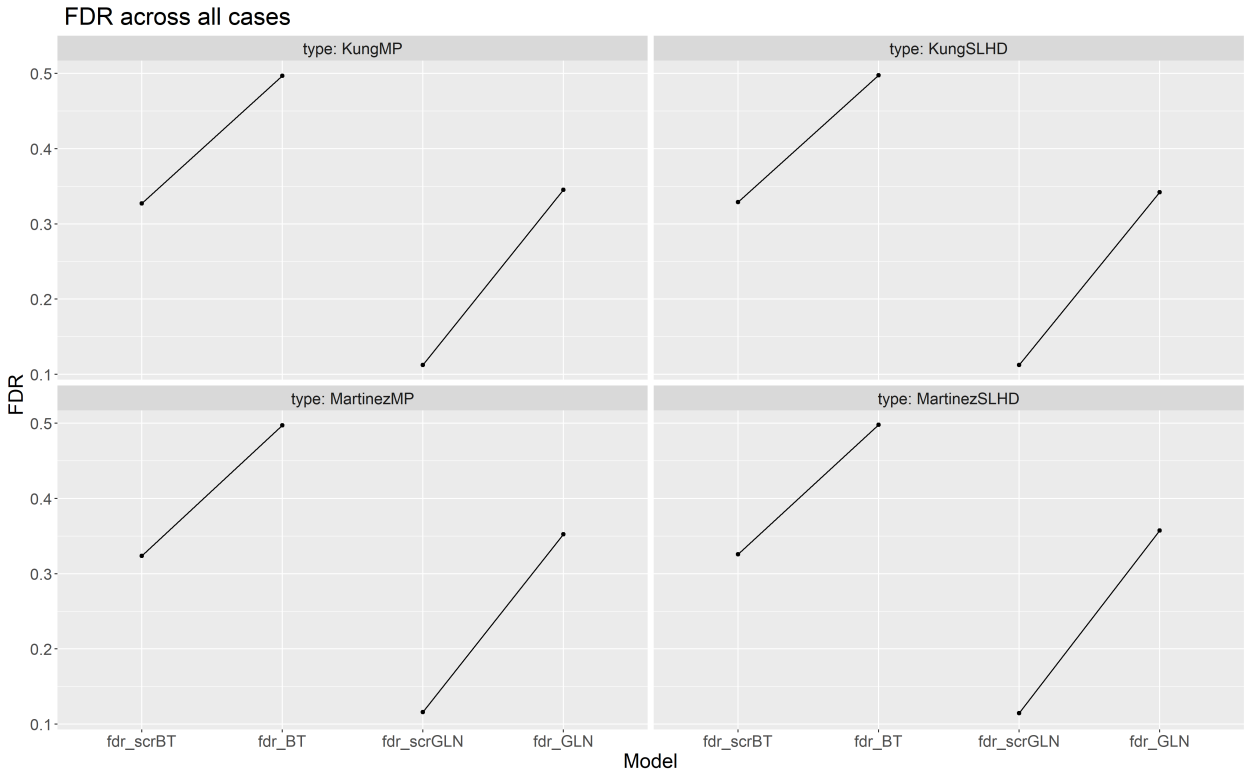


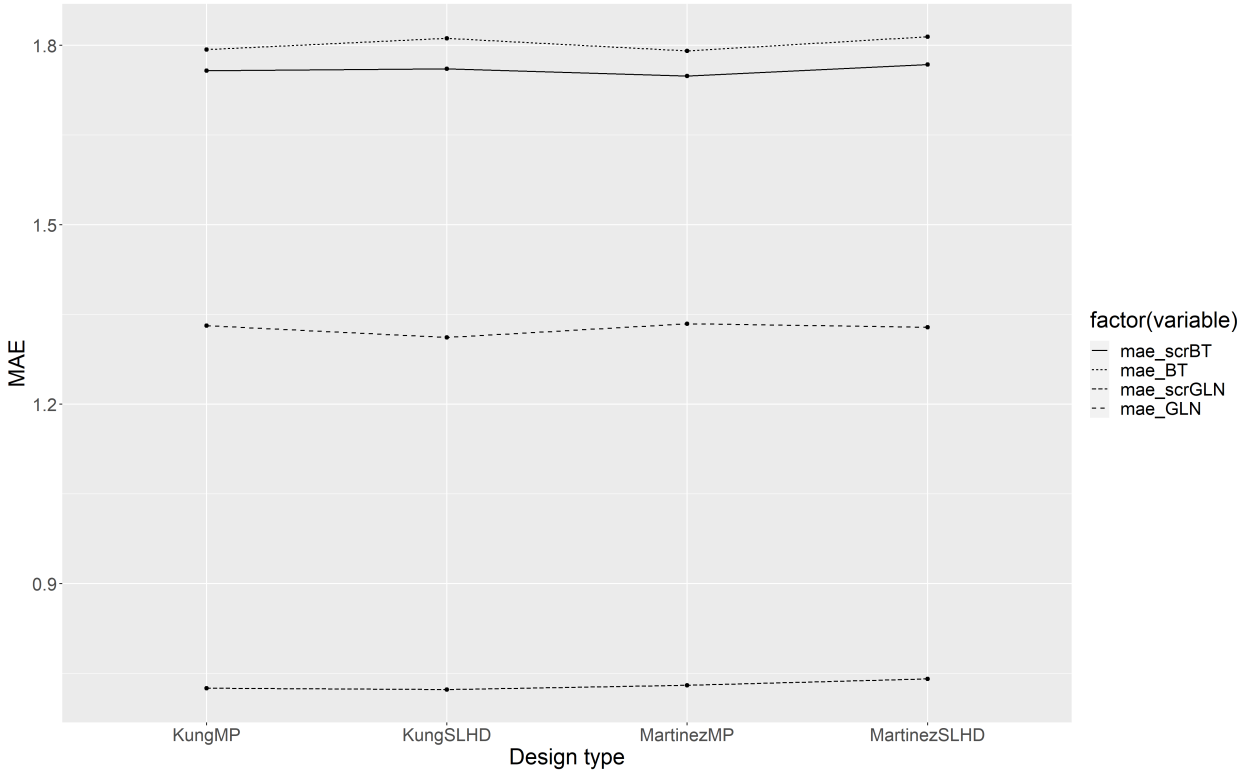
Figure 4.6: Dimension 12-False discovery rate

features.

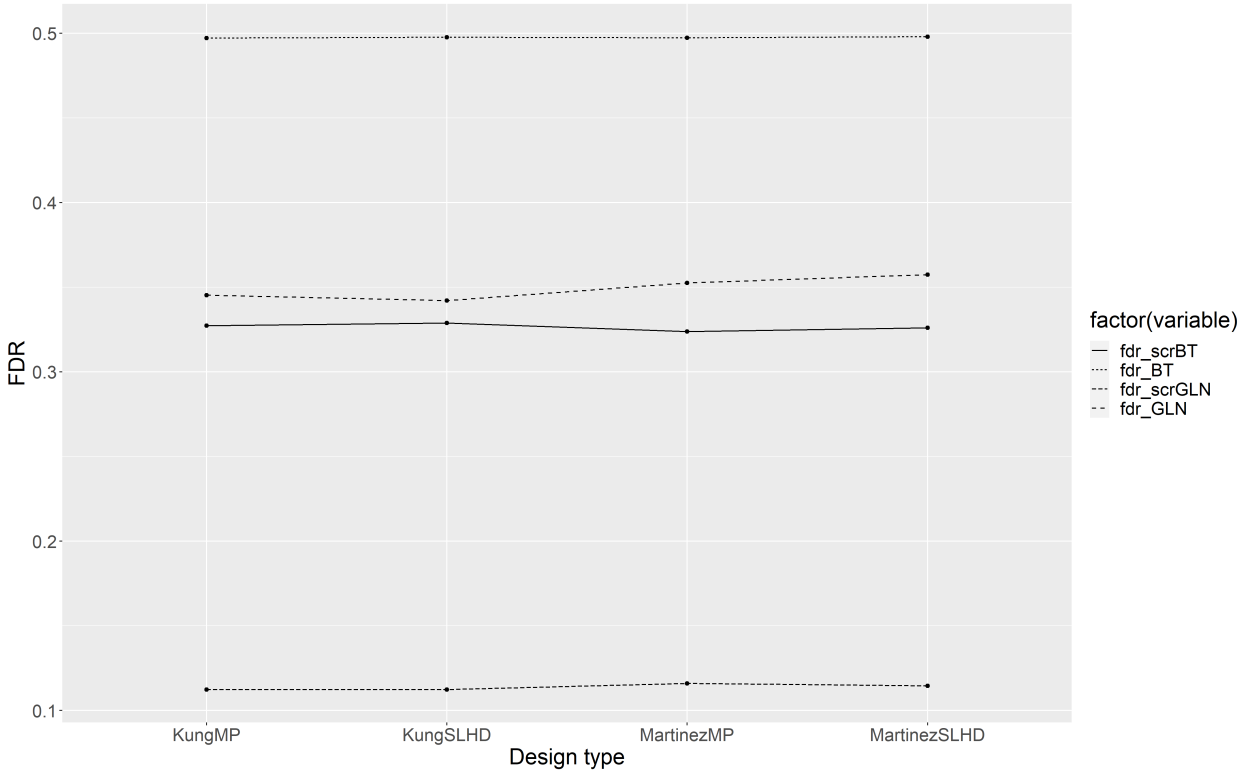
4.5 Conclusion and future work

In this paper, we proposed a screening and modeling framework for a complex mix of data structure involving mixed type of features, presence of non-linearity and interaction terms. The framework makes use of existing methodologies like MARS and Group LASSO for feature selection. From the features and model terms selected, the model is then built using Boosted Trees and Glinternet. The computational results demonstrated enhanced prediction and feature selection performance compared to the models fitted without screening process.

In our study, we modeled the ground truth including the different types of interaction terms. From the fitted models, we calculated the feature selection performance only for the main features. This work can be extended to evaluate the performance of recovering the true interaction terms as well. The challenge in identifying the true interaction terms comes



(a) Dimension 12-interaction plot(mae)



(b) Dimension 12-interaction(fdr)

Figure 4.7: Dimension 12-Interaction plots

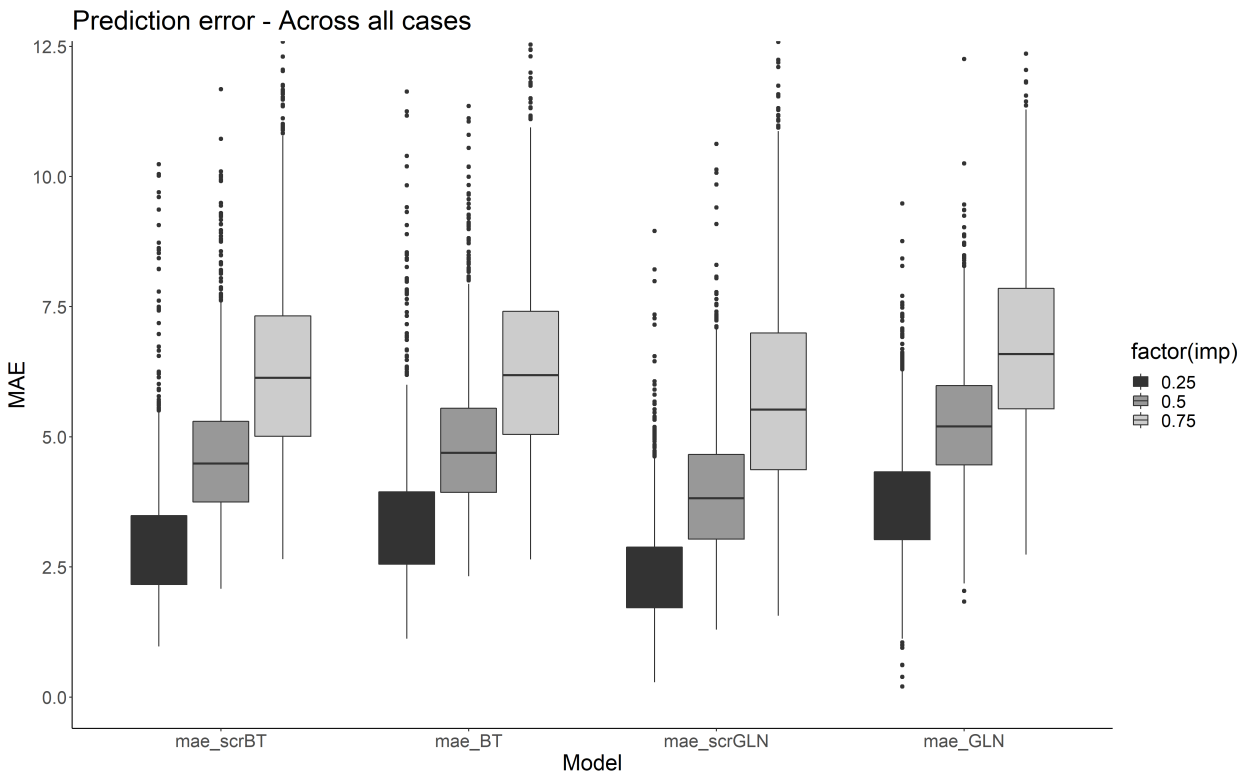
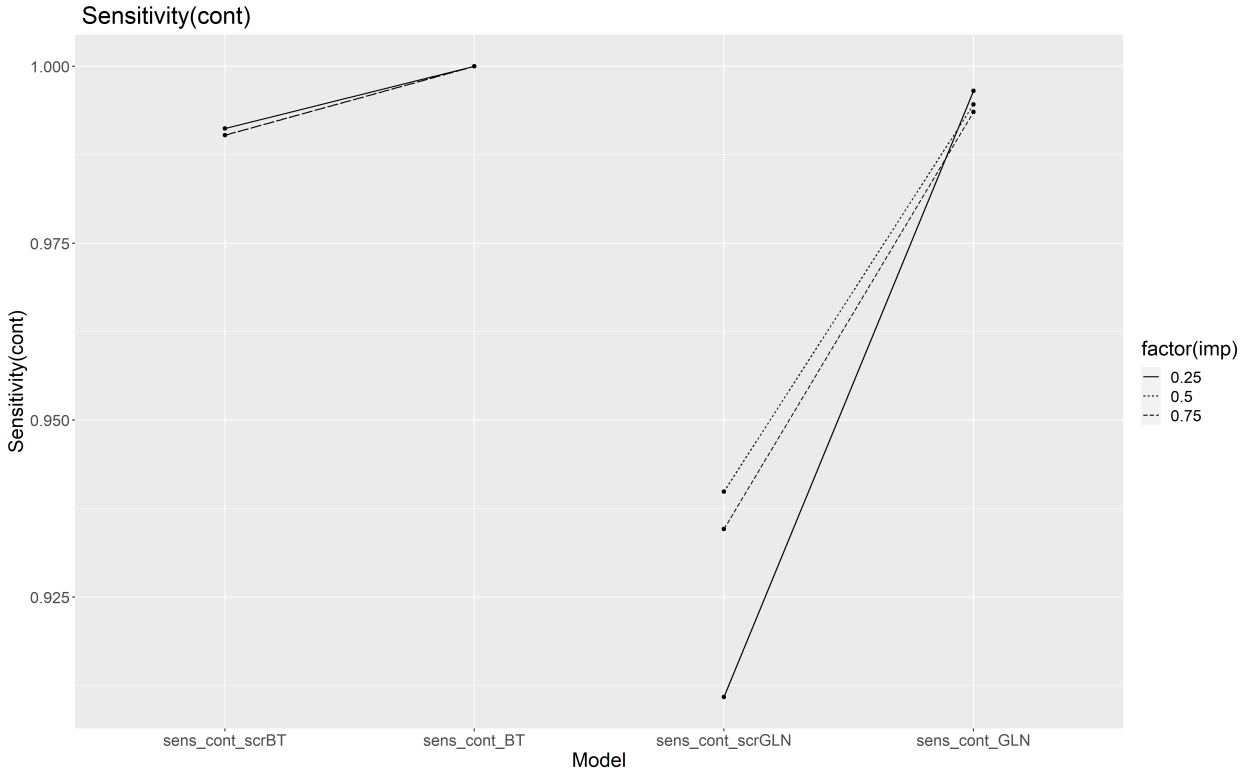
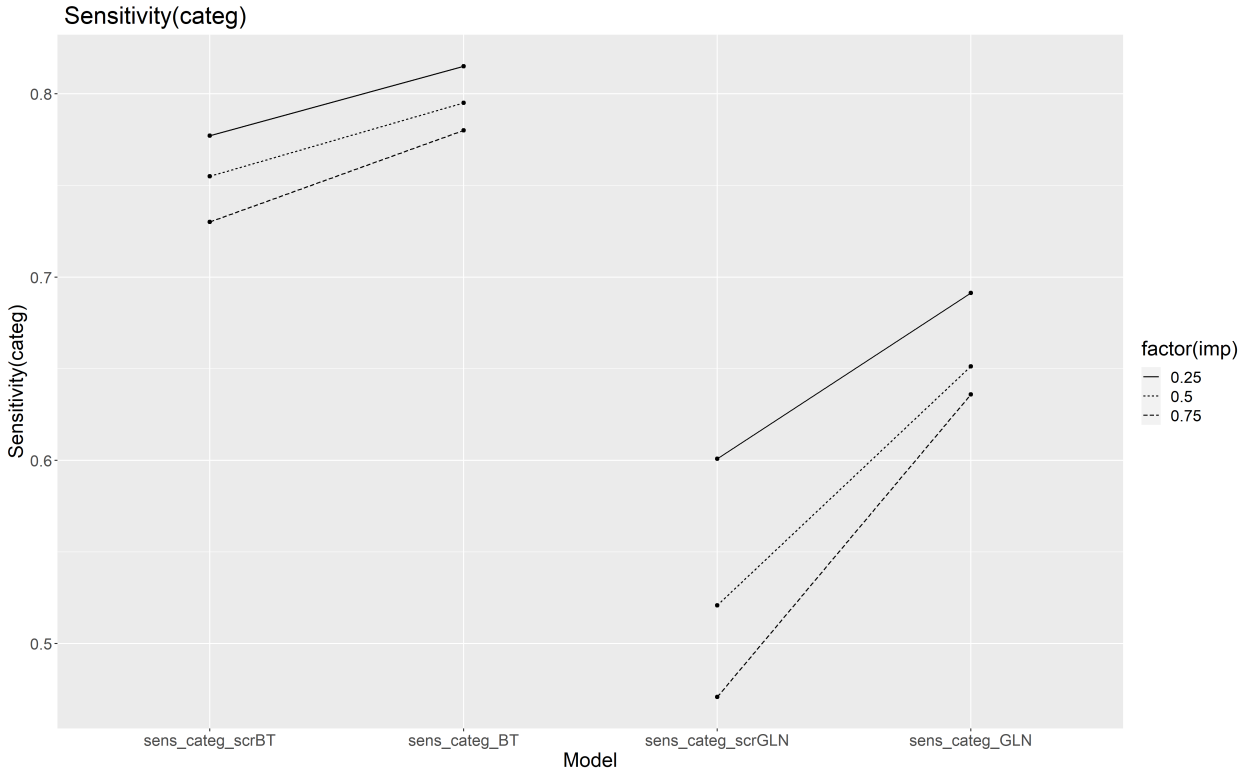


Figure 4.8: Dimension 60 - prediction error

in a boosted tree model. Gradient boosting is a tree based modeling technique, which does feature ranking as opposed to feature subset selection (as done by Glinternet). The ranking for main features is done by using a metric called relative influence. Friedman proposed a metric called 'H-statistic' to evaluate the effect of an interaction term on the model. A high value of H-statistic indicates strong interaction effects of the features considered on the response. In our studies, H-statistic did not perform consistently. Moreover, there is also the challenge to set a threshold on H-statistic as to when an interaction term can be really considered as "significant." Based on this study, there is future research area in identification of relevant interaction terms in a gradient boosting model. In addition, in our studies we considered strong hierarchy type of relationship for interactions. It would be interesting to evaluate the performance of the screening procedure on weak and anti-hierarchy types of interactions as well.

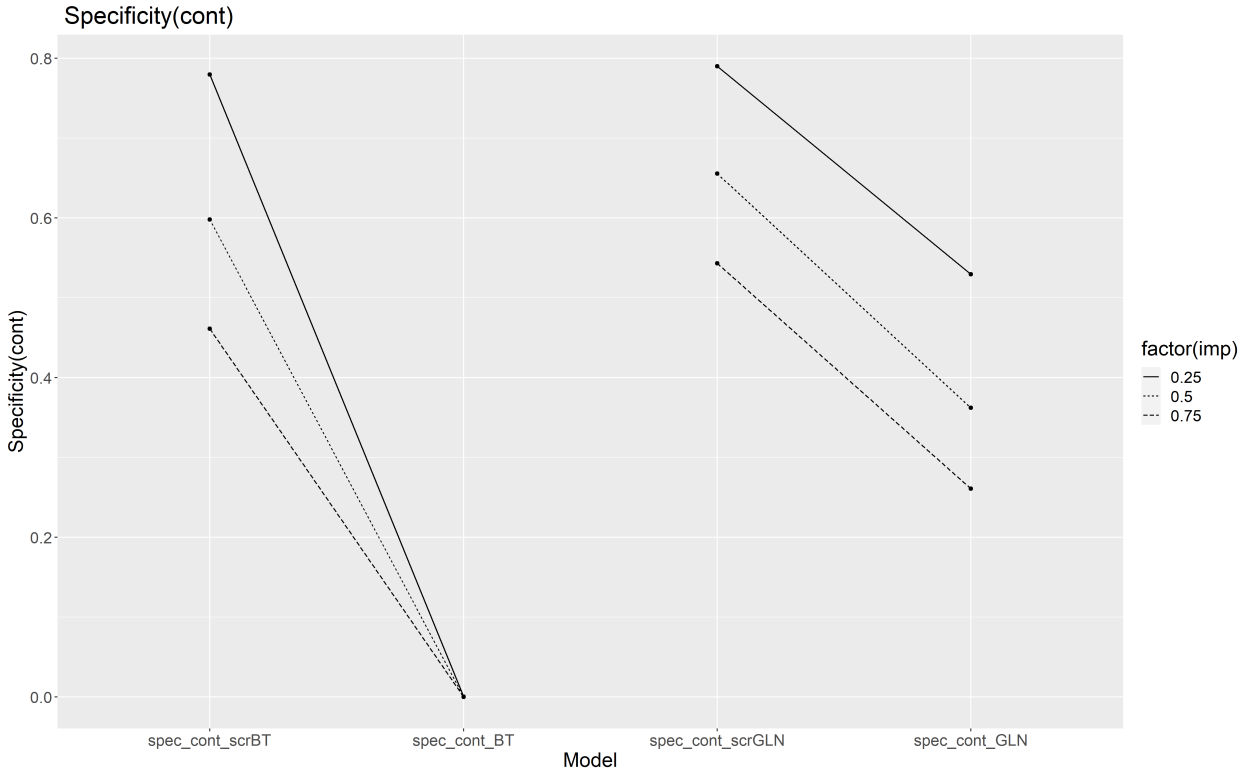


(a) Dimension 60-sensitivity(cont)

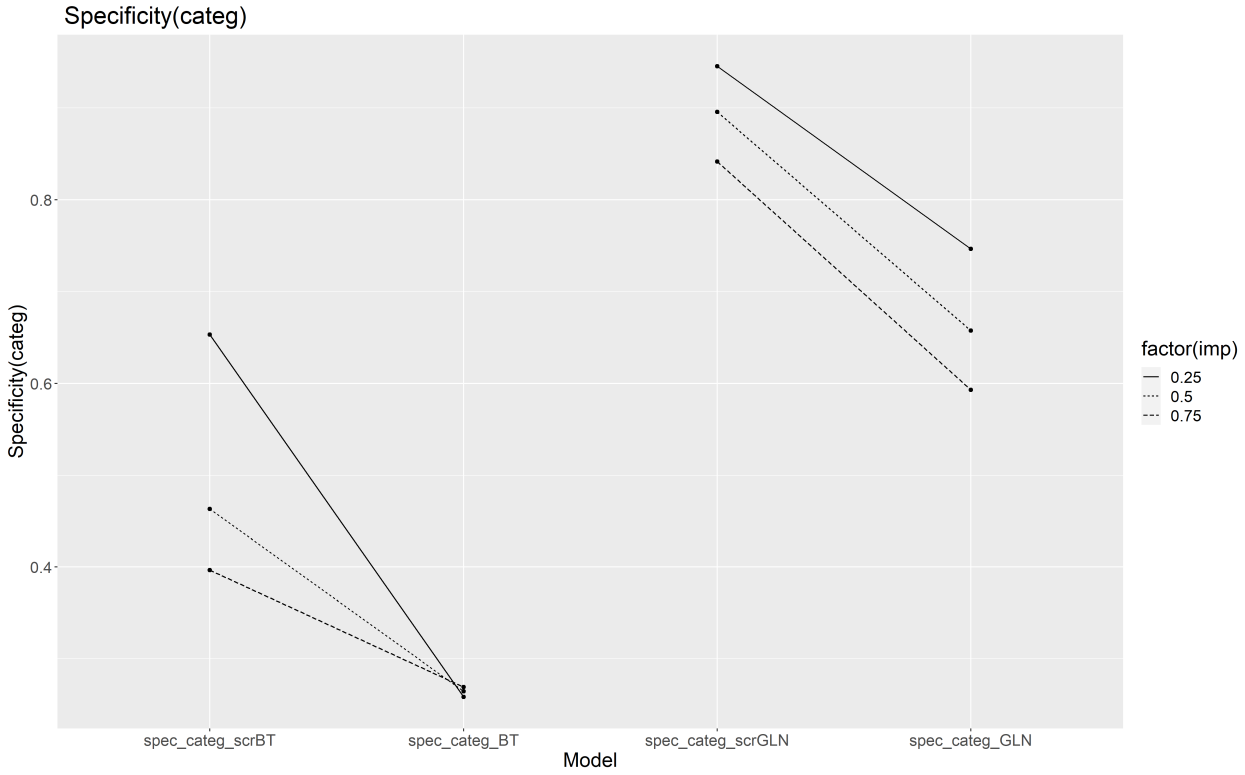


(b) Dimension 60-sensitivity(categ)

Figure 4.9: Dimension 60-Sensitivity



(a) Dimension 60-specificity(cont)



(b) Dimension 60-specificity(categ)

Figure 4.10: Dimension 60-Specificity

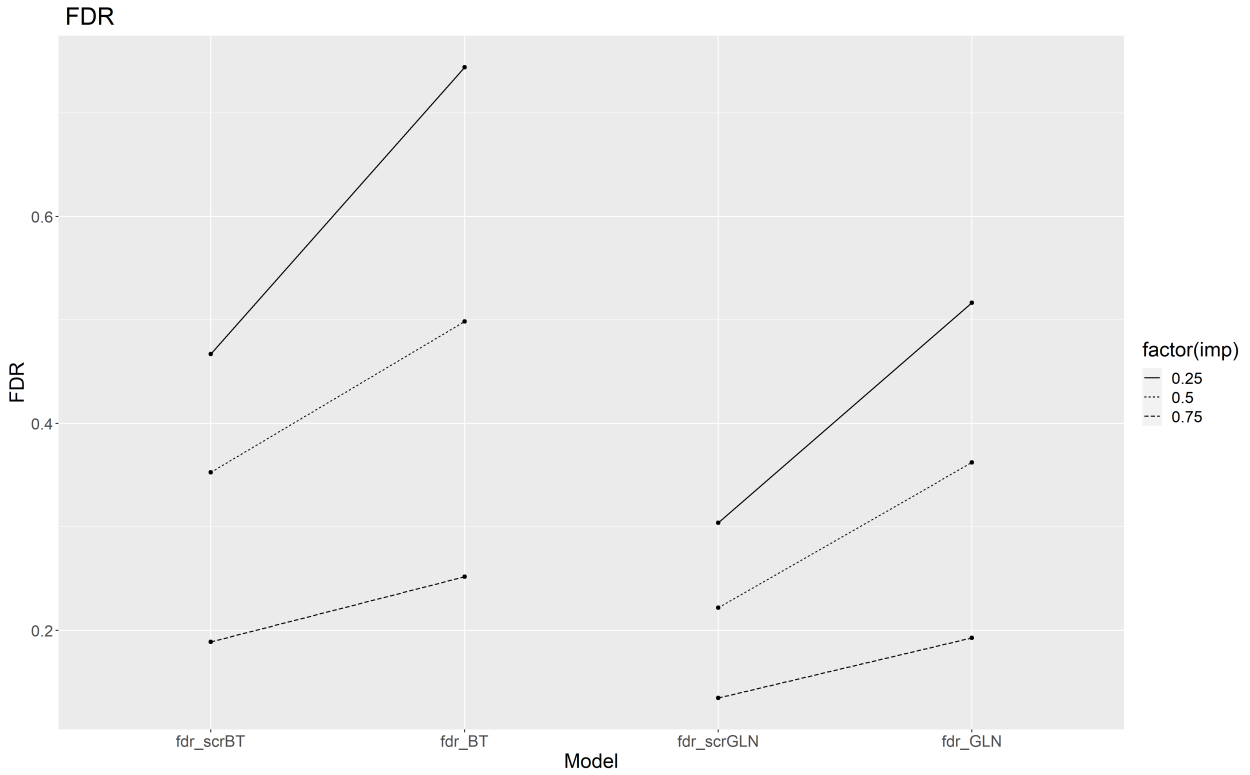


Figure 4.11: Dimension 60 - False discovery rate

5 Future work

Computer simulations are expensive in terms of time and computing power. In an earlier section, we mentioned about Ford Motor Company's simulation to take anywhere between 36 hours to 160 hours to complete. In a process where a simulation program aids decision-making systems, a long simulation time might affect the decision-making, causing operational inefficiencies. To make better use of the computer simulation programs, the concept of sequential design has been used in the past. A sequential design is initialized with a small design with few runs. Design points are then sequentially added based on certain criteria. The points that maximize information are chosen from the unexplored region are added to the design to be evaluated. In a sequential design, a 'feedback' is provided to the experimental design generation process regarding the performance of the design points already selected in the initial design. The criteria to choose new design points can be broadly classified into two categories – model (or information) based, and distance based. For only numerical fac-

tors, common distance metric like Euclidean, Manhattan and cosine distance can be used to select new design points. The maximin or minimax versions of distance calculations can also be used as a metric to ensure a good spread of points in the design space. Since this work involves many categorical factors, a distance metric capable of handling categorical factors should be used. A distance metric for numerical factors will not be meaningful on qualitative type factors. Gower's distance (Gower, 1971) is one such metric that can be utilized for this purpose. Hamming's distance (Hamming, 1950) is another commonly used distance metric used in clustering algorithms like K-means. Hamming's distance for categorical features calculates the number of instances where the corresponding levels are different for the two categorical features. There is scope to propose a distance metric for mixed factors using Hamming's and an appropriate distance measure for numerical factors. Also, a more comprehensive study involving different types of experimental design and metamodels can be done with the aim of getting a deeper understanding between experimental design-metamodel combinations. Such a study might facilitate experimenters make decisions on the type of design and metamodel for their study.

For the machine learning framework, we assumed all interactions to follow a strong hierarchy relationship. I.e. an interaction is only present if its main effects are present in the model. However, there might be situations where a feature might impact the response through interaction terms rather than as main effect. This type of relationship is termed as weak hierarchy. Hence, it might be required for these features to enter the model first as interaction, before the main effect. Research mentioned earlier can model interactions following a weak-hierarchy relationship. Glinternet, which was one of the major modeling techniques used in this work only models interactions between features whose main effects are deemed significant. A modification to Glinternet making it capable of modeling weak hierarchy interactions is an area that can be explored.

One of the major component of the proposed data-mining framework was the ability to model interactions. Form model interpretation point of view, one might be interested in

identifying significant interaction terms selected by the model. Boosted tree is an ensemble of weak learners, mainly decision trees. Being a tree-based model, boosted tree performs feature ranking instead of feature selection. Friedman defined H-statistic to evaluate the strength of interaction terms using partial dependence functions. The H-statistic is calculated between each pair of features, and higher values signify higher interaction effects. We can see that it is computationally expensive to use H-statistic. A 60-dimension dataset would include 1770 calculations. In our experiments, the values of H-statistic were unstable and the variance was high. Moreover, there is the question of how much of a high value is considered for an interaction to be significant. This leads to the potential work of proposing a metric to identify significant interactions terms for a boosted tree model. A metric based on pairs of features occurring simultaneously in splits of weak learners can be proposed. The same methodology can be extended for main effects identification. Currently boosted tree uses relative importance to rank features. The scaled relative importance is between 0 and 100, with higher values representing stronger main effects. The drawback is - if a feature is used in even one of the weak learners, the relative importance is non-zero. Based on a model, there is a potential to define a threshold, other than zero for relative importance in such a way that certain level of feature selection performance is met (like high specificity or sensitivity).

Finally, this work was motivated by a real world application in green building decision-making framework. An application like green building is appropriate for this work due to the use of computer simulation models to study a building's performance. This coupled with the presence of large number of categorical features, and potential interaction terms in the response, makes this work appropriate for such an application. There are multitude of building simulation software (EnergyPlus, eQUEST) that can simulate various aspects of a buildings performance. The Department of Energy makes available prototype buildings (DOE, n.d.) to support development of building codes. This suite of prototype buildings model covers most of the building types in The US. Using one of these prototype models as case study, this work can be demonstrated on a real-world application.

Appendices

Appendix A Results for computational studies

Link to all the results of computational studies in chapter 3 and 4: https://github.com/shirishrao1051

References

Addelman, S. (1962). Orthogonal main-effect plans for asymmetrical factorial experiments. *Technometrics*, *4*(1), 21–46.

Alam, F. M., McNaught, K. R., & Ringrose, T. J. (2004). A comparison of experimental designs in the development of a neural network simulation metamodel. *Simulation Modelling Practice and Theory*, *12*(7-8), 559–578.

ATHENA. (2017, August). *Athena impact estimator*. https://calculatelca.com/software/impact-estimator/

Ba, S., Myers, W. R., & Brenneman, W. A. (2015). Optimal Sliced Latin Hypercube Designs. *Technometrics*, *57*(4), 479–487. https://doi.org/10.1080/00401706.2014.957867

Bach, F., & Jenatton, R. (2011). *Structured Variable Selection with Sparsity-Inducing Norms Jean-Yves Audibert* (tech. rep.). http://www.jmlr.org/papers/v12/jenatton11b.html

Bach, F., Jenatton, R., Mairal, J., & Science, G. O. (2012). Structured sparsity through convex optimization. *projecteuclid.org*. https://projecteuclid.org/euclid.ss/1356098550

Bien, J., Taylor, J., & Tibshirani, R. (2013). A LASSO FOR HIERARCHICAL INTERACTIONS. *Annals of statistics*, *41*(3), 1111. https://doi.org/10.1214/13-AOS1096

Bingham, D. (2013, August). *Test functions*. http://www.sfu.ca/~ssurjano/optimization.html

Bondell, H. D., & Reich, B. J. (2006). *Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar* (tech. rep.). North Carolina State University. Dept. of Statistics.

Bose, R. C., & Bush, K. A. (1952). Orthogonal arrays of strength two and three. *The Annals of Mathematical Statistics*, *23*(4), 508–524.

Box, G. E., & Behnken, D. W. (1960). Some new three level designs for the study of quantitative variables. *Technometrics*, *2*(4), 455–475.

Box, G. E., & Wilson, K. B. (1992). On the experimental attainment of optimum conditions. In *Breakthroughs in statistics* (pp. 270–310). Springer.

Bursztyn, D., & Steinberg, D. M. (2006). Comparison of designs for computer experiments. *Journal of statistical planning and inference*, *136*(3), 1103–1119.

Candes, E., & Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, *35*(6), 2313–2351.

Chapman, W. L., Welch, W. J., Bowman, K. P., Sacks, J., & Walsh, J. E. (1994). Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *Journal of Geophysical Research: Oceans*, *99*(C1), 919–935.

Chen, V. C., Tsui, K.-L., Barton, R. R., & Meckesheimer, M. (2007). A review on design, modeling and applications of computer experiments. *https://doi.org/10.1080/07408170500232495*, *38*(4), 273–291. https://doi.org/10.1080/07408170500232495

Crombecq, K. (2011). *Surrogate modeling of computer experiments with sequential experimental design* (Doctoral dissertation). Ghent University.

De Kleine, R. D., Keoleian, G. A., & Kelly, J. C. (2011). Optimal replacement of residential air conditioning equipment to minimize energy, greenhouse gas emissions, and consumer cost in the US. *Energy Policy*. https://doi.org/10.1016/j.enpol.2011.02.065

Dean, A., Voss, D., Draguljić, D., et al. (1999). *Design and analysis of experiments* (Vol. 1). Springer.

Diwekar, U. M., Frey, H. C., & Rubin, E. S. (1992). Synthesizing optimal flowsheets: Applications to igcc system environmental control. *Industrial & engineering chemistry research*, *31*(8), 1927–1936.

Diwekar, U. M., & Kalagnanam, J. R. (1996). Robust design using an efficient sampling technique. *Computers & chemical engineering*, *20*, S389–S394.

Diwekar, U. M., & Rubin, E. S. (1994). Parameter design methodology for chemical processes using a simulator. *Industrial & engineering chemistry research*, *33*(2), 292–298.

DOE. (n.d.). Prototype buildings. Retrieved 2021, from https://www.energycodes.gov/prototype-building-models

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/J.1365-2656.2008.01390.X/FORMAT/PDF/OEBPS/PAGES/1.PAGE.XHTML

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.

Farahani, A. (2019). Uncovering underlying features for state transition modeling.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H., & Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. Sage Publications Sage CA: Thousand Oaks, CA.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*. https://doi.org/10.1007/BF01386213

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), 147–160.

Hsu, D. (2015). Identifying key variables and interactions in statistical models of building energy consumption using regularization. *Energy*, *83*, 144–155. https://doi.org/10.1016/J.ENERGY.2015.02.008

Institute, G. C. (2014, February). *Conserving modern architecture*. https://www.getty.edu/conservation/our_projects/field_projects/cmai/cmai_colloquium.html

Jacob, L., Obozinski, G., & Vert, J.-P. (2009). Group lasso with overlap and graph lasso, In *Proceedings of the 26th annual international conference on machine learning*.

Jamil, M., & Yang, X.-S. (2013). A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, *4*(2), 150–194.

JMP. (2020, August). *Fast flexible designs*. https://www.jmp.com/support/help/en/16.0/index.shtml#page/jmp/fast-flexible-filling-designs.shtml#

Johnson, M. E., Moore, L. M., & Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of statistical planning and inference*, *26*(2), 131–148.

Joseph, V. R., Gul, E., & Ba, S. (2015). Maximum projection designs for computer experiments. *Biometrika*, *102*(2), 371–380.

Joseph, V. R., Gul, E., & Ba, S. (2020). Designing computer experiments with multiple types of factors: The MaxPro approach. *Journal of Quality Technology*, *52*(4), 343–354. https://doi.org/10.1080/00224065.2019.1611351

Keoleian, G. A., Phipps, A. W., Dritz, T., & Brachfeld, D. (2004). Life cycle environmental performance and improvement of a yogurt product delivery system. *Packaging Technology and Science*, *17*(2), 85–103. https://doi.org/10.1002/pts.644

Kim, S., & Statistics, E. X. T. A. o. A. s. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *projecteuclid.org*. https://projecteuclid.org/euclid.aoas/1346418575

Kishen, K. (1942). On expressing any single degree of freedom for treatments in an \rms\rmm factorial arrangement in terms of its sets for main effects and interactions. *Sankhyā: The Indian Journal of Statistics*, 133–140.

Kohli, M., & Peralta, N. (2017). Experimental Design and Data Analysis in Computer Simulation Studies in the Behavioral Sciences. *Journal of Modern Applied Statistical Methods*, *16*(2), 3–28. https://doi.org/10.22237/jmasm/1509494520

Krige, D. (1996). A practical analysis of the effects of spatial structure and of data available and accessed, on conditional biases in ordinary kriging, In *Geostatistics wollongong*

'96. proceedings of the 5th international geostatistical congress, wollongong, nsw, australia.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, *30*(2), 195–215.

Kung, P. (2013). Multivariate modeling for a multiple stage, multiple objective green building framework.

Lekivetz, R., & Jones, B. (2015). Fast Flexible Space-Filling Designs for Nonrectangular Regions. *Quality and Reliability Engineering International*, *31*(5), 829–837. https://doi.org/10.1002/qre.1640

Lim, M., & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, *24*(3), 627–654.

Lin, Y., & Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, *34*(5), 2272–2297.

Liu, J., & Ye, J. (2010). *Moreau-Yosida Regularization for Grouped Tree Structure Learning* (tech. rep.). http://papers.nips.cc/paper/3931-moreau-yosida-regularization

Loeppky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, *51*(4), 366–376.

Lu, W. (2019, August). *Variable selection lasso.* https://www4.stat.ncsu.edu/~lu/ST7901/lecture%20notes/2019Lect6_shrink_lasso.pdf

Martinez Cepeda, D. L. (2013). Variants of adaptive regression splines (mars): Convex vs. non-convex, piecewise-linear vs. smooth and sequential algorithms.

NIST. (2009, August). *Computer aided design.* https://www.itl.nist.gov/div898/handbook/pri/section5/pri52.htm

Ortiz, F. (2012). Dealing with categorical data types in a designed experiment part i: Why you should avoiding using categorical data types. *STAT T&E Center of Excellence.*

Peter, J., & Marcelet, M. (2008). Comparison of surrogate models for turbomachinery design. *Wseas transactions on fluid mechanics*, *3*(1), 10–17.

Qian, P. Z. (2012). Sliced Latin hypercube designs. *Journal of the American Statistical Association*, *107*(497), 393–399. https://doi.org/10.1080/01621459.2011.644132

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Supplement to the Journal of the Royal Statistical Society*, *9*(1), 128–139.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, *4*(4), 409–423.

SAS. (2020, February). *Jmp*. https://www.jmp.com/en_us/home.html

Sloane, N. (2009, August). *Library of orthogonal arrays*. http://neilsloane.com/oadir/

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, *12*(1/2), 1–85.

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, *7*(4), 784–802.

Spitzley, D. V., Grande, D. E., Keoleian, G. A., & Kim, H. C. (2005). Life cycle optimization of ownership costs and emissions reduction in US vehicle retirement decisions. *Transportation Research Part D: Transport and Environment*. https://doi.org/10.1016/j.trd.2004.12.003

Tian, W. (2013). A review of sensitivity analysis methods in building energy analysis. *Renewable and Sustainable Energy Reviews*, *20*, 411–419.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

US-DOE. (2009, August). *Equest*. https://www.doe2.com/equest/

US-DOE. (2017, August). *Energy plus*. https://energyplus.net/

US-EIA. (2009, August). *Energy information admn*. https://www.eia.gov/tools/faqs/faq.php?id=86%7B%5C&%7Dt=1

USGBC. (2018, November). *Green building trends*. https://www.usgbc.org/articles/world-green-building-trends-2018-green-keeps-growing

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Wikipedia. (2009, August). *Lasso*. https://en.wikipedia.org/wiki/Lasso_(statistics)#Group_lasso

Wong, N. H. (2015). Grand Challenges in Sustainable Design and Construction. *Frontiers in Built Environment*, *1*, 22. https://doi.org/10.3389/fbuil.2015.00022

Yondo, R., Andrés, E., & Valero, E. (2018). A review on design of experiments and surrogate models in aircraft real-time and many-query aerodynamic analyses. Elsevier Ltd. https://doi.org/10.1016/j.paerosci.2017.11.003

Yuan, M., Joseph, V. R., & Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 1738–1757.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, *38*(2), 894–942.

Zhao, P., Rocha, G., & of Statistics, B. Y. T. A. o. s. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *projecteuclid.org*. https://projecteuclid.org/euclid.aos/1250515393

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.