CUNY MSDS Capstone Project

COMMERCIAL BUILDING ENERGY CONSUMPTION

ANALYSIS AND PREDICTION

March 1, 2019

John Grando
john.grando@spsmail.cuny.edu

Contents

Abstract **1**

Introduction **1**

Research **2**
Related Work . . . 2
Literature Review . . . 2

Theory and Hypothesis **2**

Data and Methods **3**
General Process . . . 3
Data Pre-Processing . . . 4

Electricity **5**
General . . . 5
Response Analysis . . . 5
Variable Selection - PCA . . . 5
Variable Selection - PLS . . . 5
Variable Selection - Random Forest . . . 6

Appendix **7**
Electricity . . . 7
Response . . . 7
PCA . . . 7
PLS . . . 8
Random Forest . . . 9

Abstract

Commercial Building Energy Consumption accounts for approximately 25%[1] of the United States energy production profile. Many economical and sociological factors are pushing owners of these buildings to reduce energy consumption and optimize performance. However, it is difficult to say whether a building is operating efficiently or not. Using publicly available data, models can be constructed to predict major fuel consumption. Keywords: building energy consumption, predicted energy consumption, baseline energy model.

Introduction

The concept of evaluating building performance typically requires two things; measuring the property in question's consumption, and comparing it to a standard practice equivalent. A baseline building consumption value is useful to inform people of expected opereational performance; however, it is only useful if it is accurate. Additionally, there are benefits to using simple, available, predictors, which can make it easier to actually use the model.

Building owners, local governments, and utility providers are all looking for ways to reduce energy consumption. Reasons for doing so can vary all the way from social responsibility to economic gain. Some people want to show off an efficient building, others want to identify propertie that are in need of improvement. however, in order to do this, a standard practice baseline value must be determined. Additionally, in order for the final product to be useful, it is important that the final set of predictors be parsimonious and be realistically available to users.

Comparing summary statistics between buildings, such as energy use per square foot, is not as simple as it seems because there are a multitude of factors that affect a building's energy consumption profile. The building use type can cause energy use to vary by a large amount; such as office buildings and refrigerated warehouses. Also, seemingly small factors, such as the hours of operation, may have significant impacts as well. This complexity of making similar comparisons creates a situation where it is difficult to determine whether a building is performing consistent with, or better than, other standard practice buildings.

Commercially, using the most popular example, ENERGY STAR [2] has implemented a benchmarking algorithm that scores buildings on a scale from 1 – 100 using market-available data. The output of this benchmarking algorithm is a unit-less score, as well as a reference 'baseline' building; however, the methodology is not released and it is unclear what factors are important to influence the energy consumption of the building. These barriers make it difficult to provide custom comparisons and nearly impossible to make batch predictions from a set of buildings, or variations of buildings.

Every few years, the U.S. Energy Information Administration (EIA) conducts a survey attempting to record pertinent features of these buildings, known officially as the Commercial Buildings Energy Consumption Survey (CBECS)[3]. While the survey is expansive (i.e. more than 600 tracked features), it is essential to create a model that is usable and only requires predictors that can easily be attained by building operators. Therefore, in this study a series of models will be evaluated

[1] EIA - `https://www.eia.gov/energyexplained/index.php?page=us_energy_commercial`

[2] ENERGY STAR - `https://www.energystar.gov/`

[3] EIA Microdata

in order to determine the most important predictors which will then be used to train a final, more complex, model. After completion of the model, the predictors will be evaluated on how easy they are to attain, and will possibly be exchanged with simpler variables that are highly correlated.

Research

Related Work

The idea of determing building energy efficiency is not a novel concept in itself. As previously mentioned, ENERGY STAR has a building benchmarking tool[4]. Additionally, the United States Green Building Council has created the Arc Platform[5] which provides benchmarking and active monitoring features. While these platforms provide building comparisons in the form of an overall score, it is difficult to explore the space around the building attribute inputs themselves as well as compare consumption of a specific building to its equivalent standard practice building. With this functionality, a more direct comparison can be made and relative environmental impact can be measured.

Literature Review

There are a variety of texts that are dedicated to the analysis of building energy consumption, and determining operating efficiency, such as ASHRAE Guildine 14. Also, there are guidelines that must be followed for buildings undergoing new construction or major renovation, which have energy compliance sections (ASHRAE Guideline 90.1, 189.1, International Energy Conservation Code). Particularly, there is a well thought out process for auditing commercial buildings, known as ASHRAE Audits, which start at the lowest level (I) and progress to the highest level (III) as the opportunity for energy and cost savings becomes more apparent [6].

Theory and Hypothesis

Commercial buildings are complex and encompass a wide variety of purposes. However, they all must be powered, and require a considerable amount of energy to operate properly. The most direct example of this complex issue is the ASHRAE energy audit process. As part of the initial audit process, an assessment of the building's overall operational efficiency is gauged. Typically, an auditor will walk through the buidling, analyze utility bills, and make the closest comparisons they can based on experience. This takes years of experience and sometimes requires highly tuned spreadsheets that have been developed over years. It can take a surprising amount of time just to determine if a building is operating efficiently or not.

The CBECS data set provide some insight as to what building attributes most greatly affect building energy consumption. Over 400 survey questions are recorded and coupled with major

[4] https://www.energystar.gov/buildings/about-us/how-can-we-help-you/benchmark-energy-use/benchmarking

[5] https://arcskoru.com/

[6] http://aea.us.org/3143-2.html

fuel consumption. These fuel sources are Electricity, Natural Gas, District Heat, and Fuel Oil. However, it would not be useful to construct a model with a large number of predictors, as it would require a large amount of time and effort to compile the necessary information in order to provide a prediction. Therefore, one of the main focuses for this study will be to extract the fewest amount of predictors necessary in order to make accurate predictions.

Given the complex nature, it is unlikely a linear regression will provide the best prediction accuracy. This point is especially highlighted by the fact the the goal of this study is produce a parsimonious set of predictors, which means a small subset must be selected. Therefore, an investigation into more complex, nonlinear, algoriithms will be performed in order to keep the number of necessary predictors as low as possible while still capturing complex interactions.

Data and Methods

General Process

Given the large number of features in the survey responses, it is not possible to analyze each one individually. Therefore, the first steps in the process will be centered around selecting a smaller subset. A few algorithms will be used in order to try and reduce bias. First, a principle component analysis will be performed. Second, a partial least squares model will be fit to the response. Third, a random forest regression tree will be used in order to try and extract any nonlinear relationships. Fourth, an attempt to construct a lasso regression model will be made. Fifth, a backward selection linear model will also be fit in order to see if an automated approach can be taken. Finally, a simple neural network model will be trained to gauge the possible effectiveness of using this model type. The magnitude and contribution percentage of each variable will be considered in selecting features from this model. Also, the various error rates from each preliminary model will be used as a benchmark for the final model performance.

After the prelminary set of models have been run and summarized, the extracted variables will be analyzed in order to verify their importance, gauge their potential predictive power, and to check whether they are easily attainable for a building operator/owner. This step is very important because it is essential worthwhile variables are used to predict the outcome. Selecting a variable that, for one reason or another, is erroneous may lead to reduced predictive power in the final model. If a variable did pass our initial analysis but doesn't actually have much predictive power (i.e. it only changes values by a slight amount) then it may not be worthwhile to select it at all. All selected variables increase the complexity of the model; therefore, we wish to only select those that will matter. Finally, the predictor must be usable, and 'knowable'. Ultimately, this tool will not be usable if a very difficult and hard to understand, and/or attain, variable value is used. These three concepts will be used in the analyzation of the candidate variables from the the preliminary analysis.

Finally, a neural network model will be built to take the verified subset of features and make predictions for the selected major fuel use. A variety of hyperparameters will be tested, using cross-validation, and compared on a common error metric. This step will reveal the optimal hyperparameter combination to use for the model. The prospective model will then be retrained on the entired entire training and validation data. This model's selected error metrics will then be compared to the preliminary models, which should be considered a floor. Next, the model's met-

rics will compared to a similarly trained neural network model that uses every available predictor, which should be considered a competitor. Once this is done, the model can then be re-assessed for feature selection as well as analyzed for the value/tradeoff of adding/removing certain features.

Data Pre-Processing

The raw data set consists of 6,720 samples and 1,119 features. However, multiple steps of pre-processing were required in order to prepare the data. Note, there are many columns which are being used as imputation flags and statistical weights which, when removed, reduced the number of features down to more tha 400. While these columns are useful to indicate where values have been imputed into the dataset by the source's own methodologies, rather than try to change back the data to the original records it has been determined that the imputed values were sufficiently applied and the dataset will not be blindly imputed any further.

After evaluating the reduced feature data set, some feature engineering efforts were taken. First, very specific cases which resulted in many null responses to follow-up survey questions (e.g. buildings less than 1,000 gross square feet), buildings open for less than a year, and features with a large amount of nulls were removed. Second, some NA entries were converted to zero when logically appropriate. For example, if a building was indicated to not be cooled, then a follow up question asking what percentage of the building is cooled was not asked, resulting in an NA. In this instance, the null value was replaced with a zero. Third, some values were removed as they simply did not apply to the study (e.g. expenditure for energy sources in USD). Fourth, nominal categorical values that had NA responses were encoded to a special value. The thinking for this approach is that if, in fact, an null value for a feature ends up being a significant predictor, then it can be analzyed what factors make this situation occur. Fifth, the categorical features were then one-hot encoded to separate columns. Finally, the response variables (e.g. electrical consumption) were normalized based on gross floor area (e.g. electrical consumption per square foot). Note, categorical and numeric features indicating the building's gross floor area have been kept in the predictor data set. The preprocessed data set was transformed to 6661 rows and 456 features (before one-hot encoding).

Electricity

General

The preprocessed data was passed to the following process in order to determine the best possible set of candidate predictors with one additional filter. Only buildings that indicated electricity being used `ELUSED` were included in the samples for this major fuel use. Then, one of each pair of predictors with correlations (above 0.75) were removed, to avoid model selection issues. Additionally, the other major fuel consumption values were removed from the set of possible predictors since separate models will be made to predict these values as well.

Response Analysis

The response data appear to be unimodal and have a heavy right skew with a median of 39.0, mean of 57.9, and max of 971.9 BTU/SF. After filtering for this model's end-use, there are 6500 samples in the data set. Due to the varying scales of all the predictors, the numeric columns have been centered and scaled before use for the non-tree regression models.

Variable Selection - PCA

The principle component analysis indicates that only 4.3% of the variance in the data can be explained in the first principle component, which then drops to 1.7% for the second principle component. These results reveal that there does not appear to be clear axes that can explain the variance of the data very well, which indicates there may be some very complex interactions taking place in the predictors. However, the top 2 predictors, based on contribution percentage to the principle components, will still be taken for further analysis:

- Variables Selected - `COOK[NO]`, `LAUNDR[NA]`
- RMSE - NA
- Rsquared - NA

Variable Selection - PLS

This model returned a promising result with an Rsquared value of 0.4318, which is pretty good considering it is a simply constrcusted model. Looking at the result, it is obvious that the use of refrigeration equipment is dominating the variable importance plot as well as some other seemingly reasonable predictors, such as the number of registers per square foot of gross floor area. The top 7 variables were selected for further review.

- Variables Selected - `RFGWinPerSf`, `RGSTRNPerSf`, `RFGICNPerSf`, `FDSEATPerSF`, `RFGCLNPerSf`, `NWKERPerSf`, `RFGOPNPerSf`
- RMSE - 53.48
- Rsquared - 0.4318

Variable Selection - Random Forest

As with the PLS, model, the resulting error metrics were promising, with slightly better RMSE and Rsquared values. The selected variables are ver similar with a few exceptions. This model has placed higher importance on a yes/no response to the presence of walk in refrigerators as well as whether or not a building is a fast food establishment

- Variables Selected - `RFGWinPerSf`, `RGSTRNPerSf`, `NWKERPerSf`, `RFGWI[YES]`, `RFGICNPerSf`, `FDSEATPerSf`, `PBAPLUS.32[Fast Food]`
- RMSE - 52.33
- Rsquared - 0.4729

Variable Selection - Lasso

Test

Variable Selection - Leaps

Test

Variable Selection - Simple Neural Network

Test

Variable Selection - Recursive Feature Elimination

Test

Variable Selection - Selected Variable Analysis

Test

Natural Gas

General

Test

District Heat

General

Test

Fuel Oil

General

Test

Neural Network Models

Electricity

Summary

Test

Performance

Test

Natural Gas

Summary

Test

Performance

Test

District Heat

Summary

Test

Performance

Test

Fuel Oil

Summary

Test

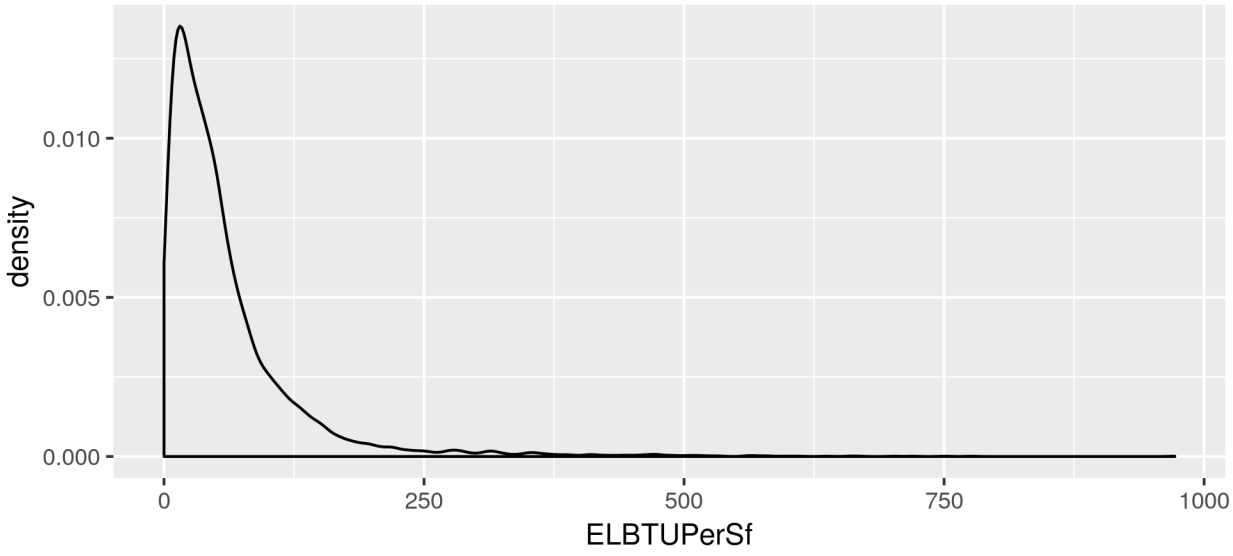
Performance

Test

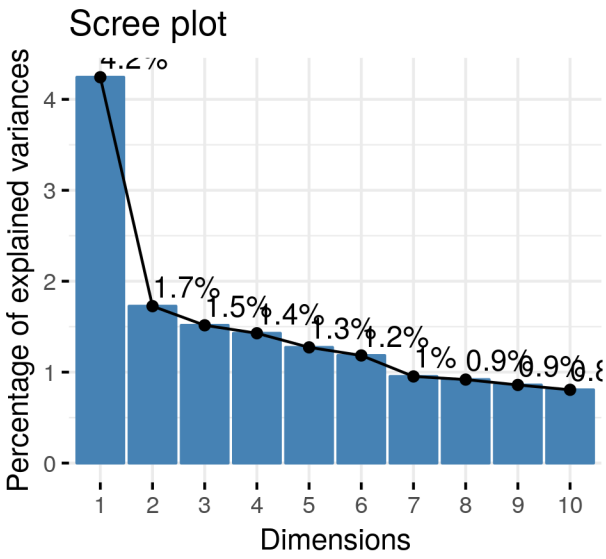
Appendix

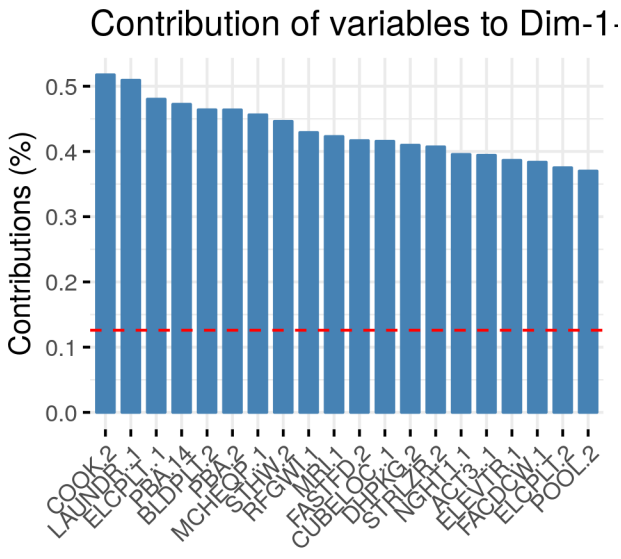
Electricity

Response

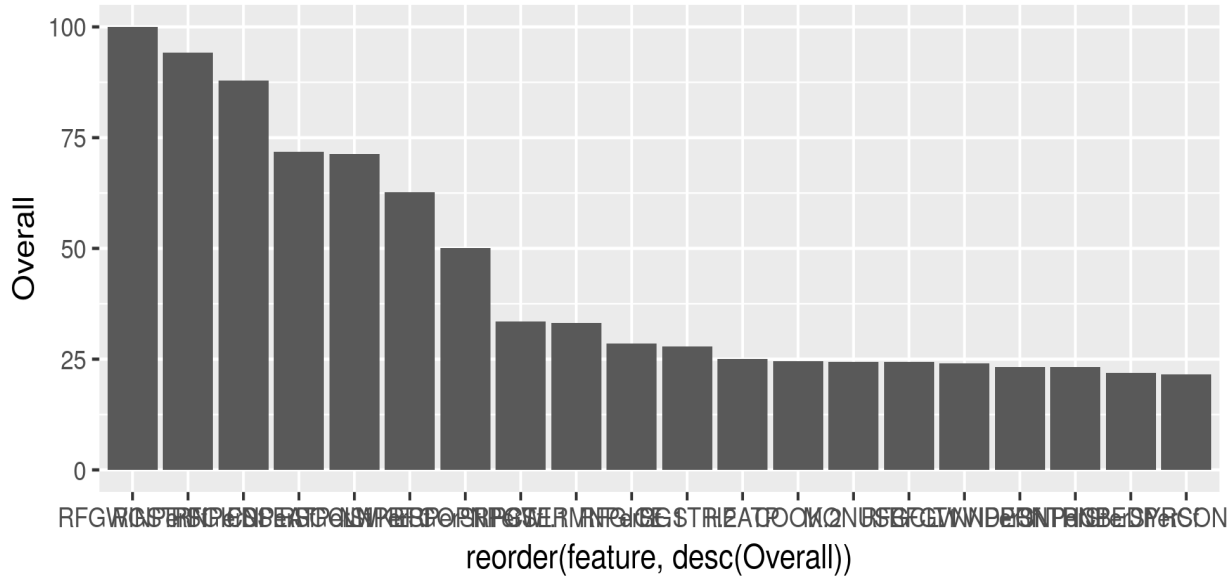


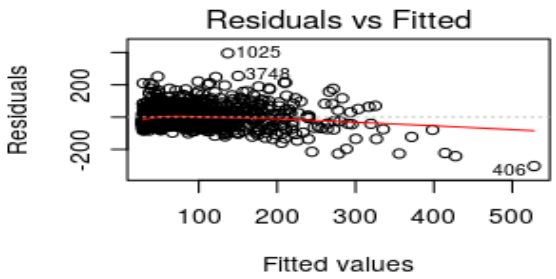
PCA

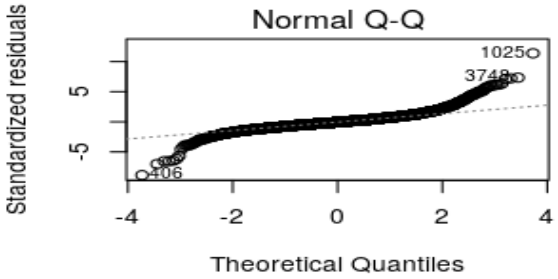


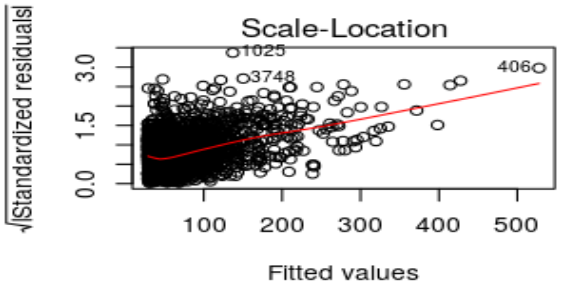


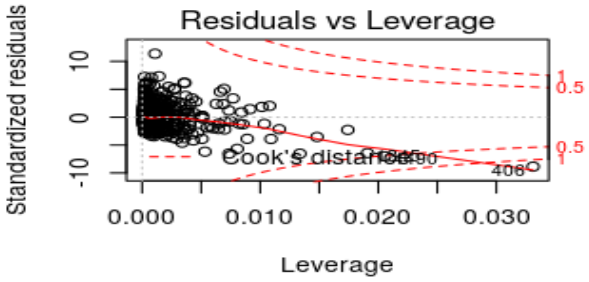
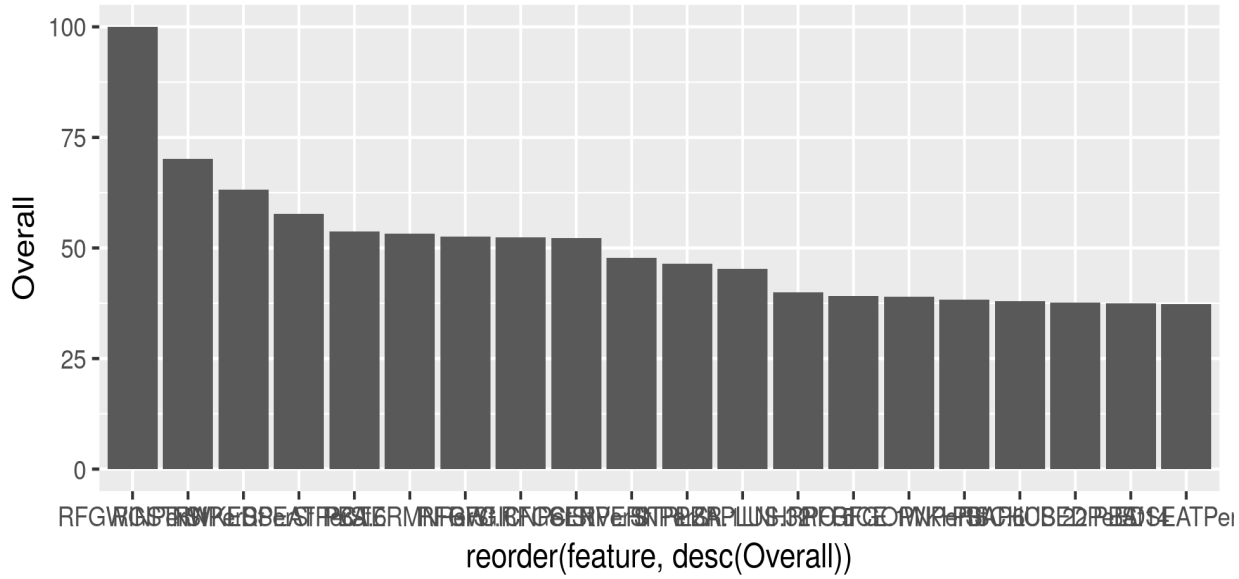
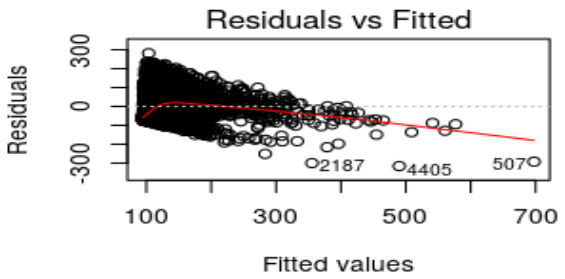
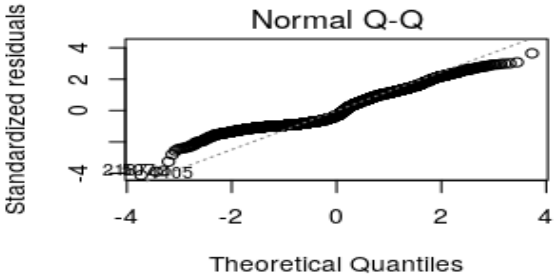
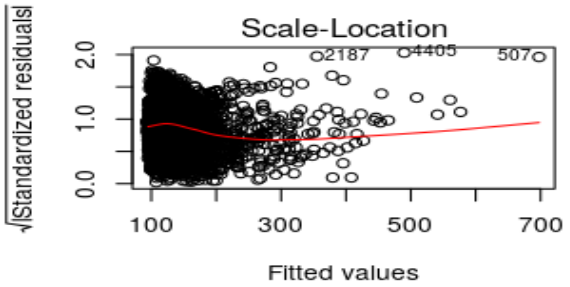
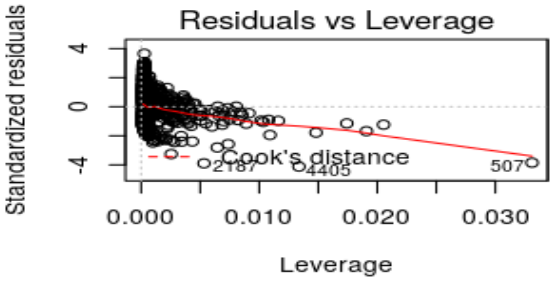
PLS


Overall
100
75
50
25
0
reorder(feature, desc(Overall))


Residuals vs Fitted
Residuals
Fitted values
1025
3748
406


Normal Q-Q
Standardized residuals
Theoretical Quantiles
1025
3748
406


Scale-Location
√|Standardized residuals|
Fitted values
1025
3748
406


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
406

Random Forest