QBUS2820 – Predictive Analytics

*Assignment 2*

Anosh S

Task A – Moneyball

# Introduction

# Predictive analytics can empower investor decision making in the real estate market with knowledge that can result in more strategic plays to maximise return. To be more specific, through investors being able to predict the price of a house based on particular characteristics, an evaluation of risks and rewards can be derived from predicted returns. Therefore, the benefits of machine learning model predictions is that it works in favour for both housing buyers and sellers.

# Typically, investor evaluations of houses have required a natural intuition and manual brute force comparison of housing characteristics to understand the value of particular houses. Thus, an individual trying to enter the real estate market can be burdened and easily intimidated by the large quantities of data available, for self-research. Hence, it is the appeal of predictive analytics in being able to automate most housing research. As a result, this should ease evaluations overall through greater efficiency, and offering potentially more reliable and accurate means to determining house prices.

# Overall, four different models were constructed, each producing varying results. The performance difference between these models were relatively large, but understandable when examining the underlying concepts of each of the models. Therefore, the predictive capabilities of the best model could be used to its greatest potential, and its benefits fully expounded upon, if the limitations of the model are understood and the appropriate domain knowledge is applied when drawing any conclusions.

# Exploratory Data Analysis

Insight into the training and test dataset, and the training set has 1570 observations while the test set has 1210 observations. Moreover, there are 80 different possible features that can be used to predict “SalePrice”.

Train dataset has 1570 rows and 81 columns.

Test dataset has 1210 rows and 80 columns.

(figure 1)

Examining into whether NA values would be a problem gave that with none being present it would not be an issue of concern. This should effectively ease the analytical process with less cleaning required.

There are a number of columns with missing values. Once the best features have been selected the observations containing the NA values will be identified and dealt with accordingly.

There are 24 columns in train dataset with missing values.

(figure 2)

Identifying the datatypes of each column, there are a number of qualitative variables that may require dummy variables to be created for these. This will be applied after an individual inspection of each of the variables in feature engineering.

MS SubClass int64

MS Zoning object

Lot Frontage float64

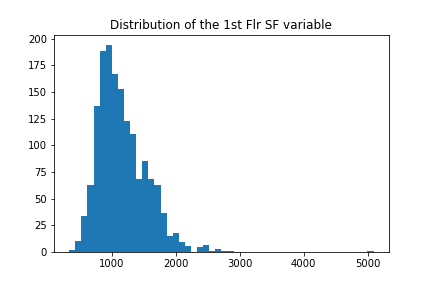
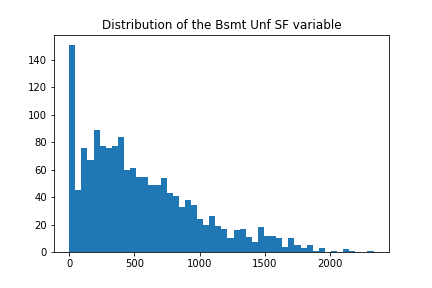
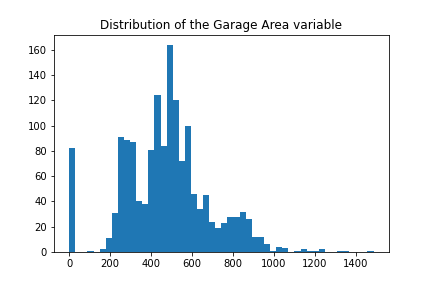
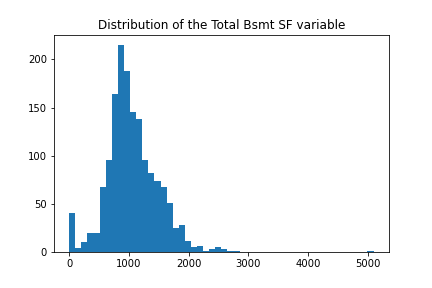
Lot Area int64

Street object

Alley object

…

(figure 3 - view full output in appendix i)

Before identifying the most ideal features, we will first delve into each of the variables to ensure that the data is logically consistent i.e. the values correspond logically with the variable that the data is trying to describe.

(figure 4 – refer to all other plots in appendix ii)

Nothing too unusual is noticeable in the data, however there is some skewness to a number of the features.

Using an IQR approach to identify outliers, these should be noted.

The dataset has 966 outliers.

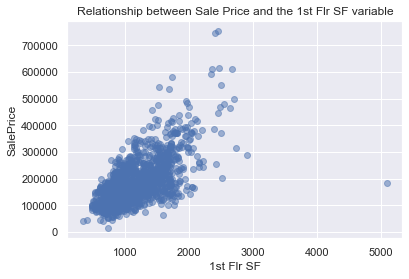
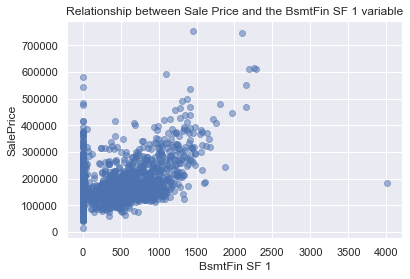
(figure 5)

Since, there were 966 observations that contain at least one outlier under one of its features, this represents a significant proportion of the train dataset and thus were not ignored in training.

# Feature Engineering

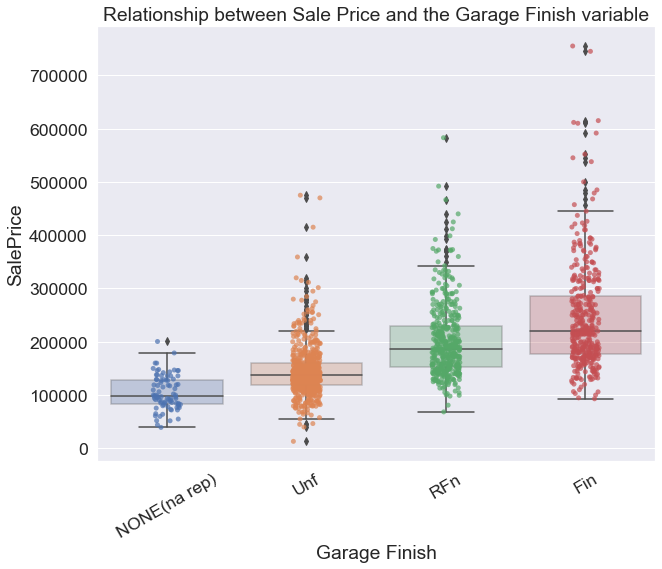
First, we will examine for the most relevant numerical variables in the dataset. Later, the best selected categorical features will be converted to dummies so that they can be usable for predictive modelling.

Subtracting the year it was sold in to the year it was constructed the new variable ‘Age’ was created. Also, subtracting the last year it was remodeled or had additions made to the house to the year it was sold was added as the “Last remod/add” feature.

The following is a series of scatter plots of the numerical features against “SalePrice” to understand what relationships may be present between these variables.

(figure 6 – refer to all other plots in appendix iii)

Viewing the scatter plots the best variables that were considered were "Gr Liv Area", "Garage Area", "Age", "Last remod/add", "Total Bsmt SF", "1st Flr SF", '2nd Flr SF', “BsmtFin SF 1”, “Full Bath”, “TotRms AbvGrd”, “Fireplaces” and “Garage Cars”. This is due to these showing some form of a non-linear relationship.

To see if any relationship is present between the categorical variabels and the “SalesPrice” predictor, box plots were made between these.

(figure 7 – refer to all other plots in appendix iv)

The best categorical features would be "Kitchen Qual", "Garage Finish" and "Overall Quality". This was on the basis that there was a sufficient number of observations between the different categories, a clear pattern was evident and that the variation (the whiskers of the boxplots) between groups did not overlap excessively. "Alley" would have also been considered, however, the presence of NA values that did not show a pattern or a distinct variation between the other categories meant NA values had little meaning.

Before confirming the final subset of features, to ensure that there were no issues in training and evaluation the features that contained missing values were to be dropped or imputed with an average.

The missing values under "Total Bsmt SF" and "BsmtFin SF 1" were replaced with their respective mean, since only one observation was found to contain these missing values under the features. Moreover, provided that only one observation in the test set had an NA value for "Garage Area" and "Garage Cars" it was best to impute the values of this single row with their means rather than removing the entire "Garage Area" and “Garage Cars” variables. Also, it was noticed that during model evaluation some of the observations in the test dataset under "Kitchen Qual" contained observations that had a category not defined in the training set. Hence, “Kitchen Qual” was disregarded from both sets.

Therefore, the following features that were considered for model construction and evaluation were "SalePrice", "Gr Liv Area", "Garage Area", "Age", "Last remod/add", "Total Bsmt SF", "1st Flr SF", "Kitchen Qual", "Overall Qual", "Garage Finish", '2nd Flr SF', 'BsmtFin SF 1', 'Full Bath', 'TotRms AbvGrd', 'Fireplaces' and 'Garage Cars'.

To finish feature engineering the train and test narrowed dataframes were finalised with its numerical variables being standardised. Typically the categorical variables would have been dummified here too, however, to simplify the model construction workflow this was done just before the subset of features were used to train the model.

# Methodology and Modelling

## Multiple Linear Regression

Having derived the most correlated variables with 'SALARY', we ensure to obtain the most optimal model not through simply using the selected features but by an assessment with each combination that can be made with the predictors.

Also, since multicollinearity is present between "WS", "DWS" and "OWS" we want to create combinations that only select one of these each time a set of combinations is made to be used in the model. This way we can ensure each combination does not suffer from multicollinearity.

First presence of outliers will be addressed, through identification and then their subsequent removal. This is critical since the influence of outliers can have a large impact on predictor performance, this will be shown below where the performance differences in the cross validation and test sets are apparent.

-------------------------------------------

---------------WITH OUTLIERS---------------

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best mlr model uses the features ['PER', 'DWS', 'Minutes', 'USG']

CV RMSE: 3.7293

Test RMSE: 4.1146

…

-------------WITHOUT OUTLIERS--------------

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best mlr model uses the features ['WS', 'Minutes', 'USG']

CV RMSE: 3.4782

Test RMSE: 4.0660

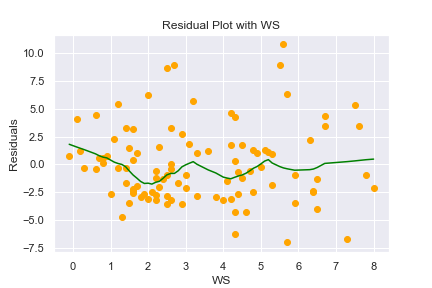
(figure 10 – full results in appendix v)

Therefore, comparing the cross validation and test performance between the dataset with and without outliers, the difference is clear. Since the model without outliers provides the most optimal results, we will continue to use the best model that used training data which excluded the outliers.

Now, considering the performance of each of the model possibilities without outliers the best one used features “WS”, “Minutes” and “USG”. This was selected as the most desirable, as it achieved the lowest cross-fold validation RMSE of $3.4782 million while also attaining an above target test RMSE of $4.0660 million. Although, there were models that performed better with the test dataset it is not correct methodology to select the best model on this basis.

While the model did perform overall well, it should be understood that multiple linear regression has a few assumptions that may impact its validity as a useful model. Therefore, these will be assessed to ensure that we are aware of limitations.

The first assumption is that the features used follow some linear relationship. As observed in the pair plots and a thorough assessment of the correlations earlier, the selected features satisfy this assumption, as it shows some linearity with “SALARY”.

Next, we need to assess for exogeneity i.e. the residuals are independent and uncorrelated. Since we would need to determine for omitted variable bias (OVB) through assessing the correlations of each variable, the fact that the dataset does not contain all the official NBA metrics of each individual, it is not possible to use OVB in understanding the assumption for exogeneity. However, through a residual plot with each of the predictor variables we should be able to determine if there is an indication of possible correlations amongst the errors.

(figure 11)

Examining the plots above, it is apparent that the residuals relatively random and thus suggest that the errors are not correlated. While we could accept this assumption as satisfied it is important to consider that our lack to be able to assess for OVB limits our confidence to do so and must be acknowledged.

Moreover, the residual plots also imply the presence of homoscedastic errors. This means that the errors are of constant variance, as indicated above. Hence, the assumption for homoscedasticity is satisfied.

In assessing for finite fourth moments, it can be concluded as satisfied since the predictor and response variables can only exist as finite values.

A simple VIF assessment of the selected features should be able to delineate if multicollinearity is present and thus if the assumption for no perfect collinearity is satisfied.

VIFs with WS, Minutes, USG: [2.0663, 2.1004, 1.0888],max\_vif: 2.1004

(figure 12)

Since, the features all possess a VIF less than 3, its highest being only 2.1, it is appropriate to conclude that the assumption for no perfect collinearity is satisfied.

Unfortunately, the assumption for independence of the predictors might be problematic, given that the performance of an individual is hard to assess since much of it depends on overall team performance. Therefore, there is the slight possibility that some of the observations may be dependent across players from the same team.

Therefore, the use of the multiple linear regression model is relatively justified given its overall good performance and satisfying of most assumptions.

## kNN Regression

Similarly, to select the most optimal model we will construct one for every different combination of the selected features and then assess each combination against each other with its best k (i.e. the nearest neighbours parameter).

Since, we want to be sure that the models are not subject to multicollinearity, as we did earlier, we will attempt to build a model for each of the multicollinear features "WS", "DWS" and "OWS". That means each set of combinations will have a multicollinear feature replaced with a different one.

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best kNN model uses the features ['PER', 'DWS', 'Minutes', 'USG']

CV RMSE: 3.6202

Test RMSE: 4.0645

Number of neighbours: 11

(figure 13 – full results in appendix vi)

After formulating each of the different model possibilities the best model used features “PER”, “DWS”, “Minutes” and “USG” with the k set to 11. This was decided as the most optimal, since it achieved the lowest cross-fold validation RMSE of $3.6202 million. It also successfully achieved an above target test RMSE of $4.0645 million, however, the model using the features “PER”, “WS” and “Minutes” was able to attain a lower test RMSE. Having said that, it is not correct methodology to select the best model on the basis of test data performance, but rather cross-validation performance. This is why we have accepted the model with “PER”, “DWS”, “Minutes” and “USG” as the best.

Since, the kNN model is non-parametric there are no assumptions made on the underlying data distribution, it is possible to accept the above best model as an alternative means to predicting “SALARY”.

## Support Vector Regression

Support vector regression (SVR) utilises similar principles as it does in support vector machine classification and may be a possible model solution alternative. SVRs operate on the basis of kernels, where they can adapt to either linear or non-linear type data and may offer more flexibility and control (Learning and Sethi, 2020). Moreover, the model is more robust to outliers and is less computationally expensive as compared to kNN and MLR models (Regression, 2020). It is through these advantages that the model could possibly outperform the other two.

Utilising principles in support vector machines, support vector regression selects the hyperplane between the features that covers the most points given a margin of tolerance from the hyperplane. SVR as a quadratic programming model, it uses a different method to optimising for errors in training, and that is through only taking the distances of points that are not bound within the constraints i.e. the margin of tolerance (Learning and Sethi, 2020).

As observed earlier in the scatter plots between the features and "SALARY", the best relationship that could be observed amongst the variables was a linear association. Therefore, we will use a linear kernel for the SVR model.

Although, a linear relationship is formulated amongst the features, no comparison of performance between training datasets, with and without outliers, is necessary. This is due to SVR being robust to outliers as mentioned before.

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best svr model uses the features ['PER', 'DWS', 'Minutes', 'USG']

CV RMSE: 3.7372

Test RMSE: 4.0768

(figure 14 – full results in appendix vii)

The best model used the features 'PER', 'DWS', 'Minutes' and 'USG', attaining a CV RMSE of $3.7372 million and a test RMSE of $4.0768 million.

Similarly, to kNN regression, SVR does not require any assumptions to be satisfied. Thus, the above best model would be a valid approach to predicting "SALARY" of NBA players on the basis of 'PER', 'DWS', 'Minutes' and 'USG'.

# Model Evaluation

|  |  |  |
| --- | --- | --- |
|  | CV RMSE | Test RMSE |
| MLR | 3.4782 | 4.0660 |
| kNN | 3.6202 | 4.0645 |
| SVR | 3.7372 | 4.0768 |

(figure 15)

Having performed model construction, the following delves into the ability for the each of the best models to generalise with unseen data i.e. test data. Overall, each of the best models of each of the types, achieved an above target test RMSE of $4.1 million. Also, it should be noted that the bias-variance tradeoff is addressed most optimally with the use of cross validation. These are good results, however, the relatively large difference between the cross validation and test performance across all the models brings some concerns. One way of possibility as to why this was the case is the fact that the split of data between the training and test sets does not use the conventional 80 to 20, but rather a 50 to 50 split. Hence, the models could have generalised better if it had a larger training set that made them less prone to unfamiliar observations that inflate the test RMSE.

The multiple linear regression model achieves the best performance overall, attaining the lowest cross validation RMSE and close to lowest test RMSE. The kNN model does slightly worse in cross validation, however, it is important to recall that MLR produced only the best model once outliers were excluded, while kNN achieved above target results even with outliers. Therefore, the higher cross validation RMSE can be accounted to these outliers and also the tendency of kNN models to be typically sensitive to noisy data. Otherwise, its achieving of an above target RMSE is indicative of good generalising ability. Also, since the kNN model is non-parametric and does not make any assumptions on the distribution of the data and the fact that it performed similar, if not, just as well as MLR, it is a testament to kNN being an effective alternative to predicting NBA salaries.

It was expected that the SVR model would be robust to outliers and thus was trained with the data that included them. This robustness is evident in the results above. Its cross validation RMSE is still relatively worse than the other two, however, it was still able to achieve above target and thus implies that it can generalise well. Thus, it is recommended that further formulation testing that adjusts to different values for the margin of tolerance could potentially allow for us to derive a more improved model.

Therefore, the kNN model is the best performer overall, since the MLR’s validity is questionable given some of its assumptions could not be satisfied confidently and that SVR achieved a slightly worse performance in terms of cross validation and test RMSE.

# Conclusion

In conclusion, the outcomes of the research in predictive modelling of NBA salaries are quite promising. Each of the models achieved above target test RMSEs and gave results that were consistent with what was expected. While these reasons support the practical implementation of these models in assisting NBA teams, there are limitations that need to be noted. Improvements to consider is that the data provided could have been organised such that the training and test split was more conventional, so a better generalisation performance could have been achieved. Moreover, the data could have been more complete with a wider selection of metrics on players, so that they could have expanded the predictive abilities of the models, through a more accurate representation of NBA players overall. Also, while the predictors selected were posited to have some correlation with salary it is important to recognise that these are associations with salary and not causal relationships. Given these limitations, important considerations for future research is that the provided data is more complete with a better train-test split, the SVR model is reformulated against different margins of tolerance to potentially derive a more improved model, and that other models are experimented to explore other potentially better alternatives.

Therefore, while these models are definitely useful tools, these should not be the sole reasons to any final decisions. This requires that the users of these tools makes use of domain knowledge and takes into account other factors that may not be considered by them, such as the constantly changing circumstances of salary conditions in the NBA league.

Task B – Gradient Ascent

# Exploratory Data Analysis

The column and row dimensions of the dataset are as follows.

The dataset has 506 rows and 14 columns.

(figure 16)

Determining the number of NA values, none were present.

There are 0 columns in train dataset with missing values.

(figure 17)

Each of the features are of numerical datatypes with only one binary categorical variable, that being “chas”, as reported on the data description document and can be confirmed in the distribution plots (refer to appendix ix).

crim float64

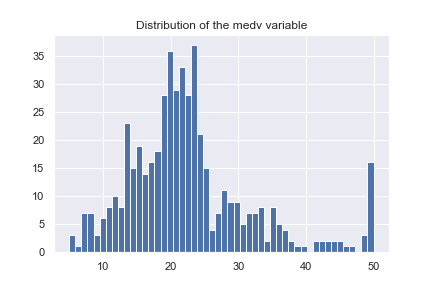
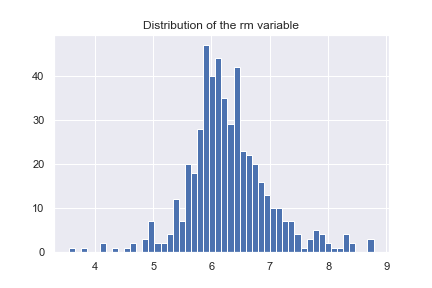
zn float64

indus float64

chas int64

…

(figure 18 - view full output in appendix viii)

Ensuring that all inputted values of the dataset are logically consistent, the distribution plots below will give insight.

(figure 19 – refer to all other plots in appendix ix)

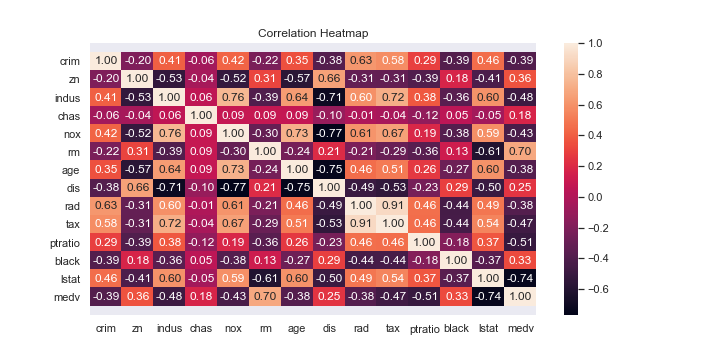
Viewing the distributions and comparing them to the data descriptions provided, the variables appear to be logically consistent.

In detecting for outliers, using the IQR method, 288 were found.

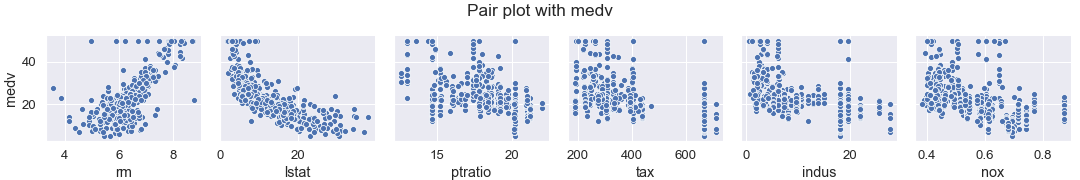
The dataset has 288 outliers.

(figure 20)

# Feature Engineering

To derive a good feature set for linear regression, we will conduct dimensionality reduction using a correlation heatmap to select the most relevant features.

(figure 21)

Referring to the above plot, the top 3 most correlated features with “medv” are “rm”, “lstat” and “ptratio”. “rm” is positively correlated while the other two are correlated negatively. To confirm the strength and linearity of these relationships a pair plot will be examined. The next three most correlated features will also be included in the pair plot to see if these variables may be better at predicting "medv".

(figure 22 – for complete pair plot refer to appendix x)

The pair plot shows that “rm” has a relatively strong linear relationship with “medv”, while “lstat” seems to follow a rough exponential form. However, “lstat” would remain to be an appropriate feature to explain “medv” in the linear model. The rest seem to share similar patterns in a relationship with “medv”, and so on the basis of attaining the highest correlation, “ptratio” remains as the third best feature to include.

It should be noted however that, “lstat” and “ptratio” share a -0.61 correlation and thus introduces a concern for possible multicollinearity. This will be assessed further with the VIF.

VIFs with "rm", "lstat", "ptratio": [1.6534, 1.6794, 1.1981], \max\_vif: 1.6794

(figure 23)

The results show that no multicollinearity is present between the best three features. Therefore, it is possible to conclude that “rm”, “lstat” and “ptratio” are the best to use in a linear regression model to predict “medv”.

# Gradient Ascent Implementation

Having constructed the gradient ascent algorithm, the algorithm is run through a leave one out cross validation across different values of alpha, whilst keeping the number of iterations fixed. This is because, it is expected that the performance of gradient ascent increases with greater values in the number of iterations. Thus, in assessing for different values of alpha whilst keeping the number of iterations fixed would be a sufficient method to deriving the maximum likelihood. Therefore, we fixed this value at 10 so to ensure algorithm had iterated enough to reach the optimum whilst retaining computational efficiency.

------------------|-------------------

ALPHA | CV RMSE

------------------|-------------------

0.001 | 0.9925

0.01 | 0.9322

0.1 | 0.7035

0.2 | 0.6743

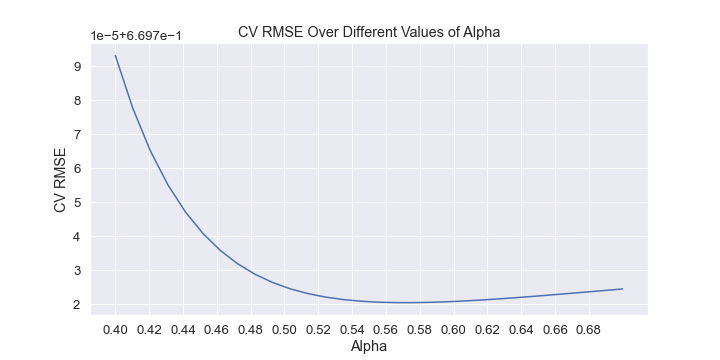
0.5 | 0.6697

1.0 | 0.6702

3.0 | 796553.9641

6.0 | 2572815658.9413

(figure 24)

Considering the provided output, the best alpha was 0.5, since it achieved the lowest cross validation RMSE. Now, it is possible to find a more precise value for alpha through running a cross validation over alphas close to 0.5.

(figure 25)

Therefore, from viewing the plot, the most optimal value of alpha is 0.57. This is since cross validation RMSE is depicted to be the lowest at this point as shown above.

References

Academy, U., 2020. Determinants Of NBA Player Salaries. [online] The Sport Journal. Available at: <https://thesportjournal.org/article/determinants-of-nba-player-salaries/> [Accessed 13 October 2020].

KNN, W. and Jones, D., 2020. Why Do You Need To Scale Data In KNN. [online] Cross Validated. Available at: <https://stats.stackexchange.com/questions/287425/why-do-you-need-to-scale-data-in-knn> [Accessed 10 October 2020].

Learning, S. and Sethi, A., 2020. Support Vector Regression In Machine Learning. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/> [Accessed 14 October 2020].

Regression, S., 2020. Support Vector Regression | Learn The Working And Advantages Of SVR. [online] EDUCBA. Available at: <https://www.educba.com/support-vector-regression/> [Accessed 14 October 2020].

Scikit-learn.org. 2020. 1.4. Support Vector Machines — Scikit-Learn 0.18.2 Documentation. [online] Available at: <https://scikit-learn.org/0.18/modules/svm.html#svm> [Accessed 13 October 2020].

# Appendix

## Appendix i

MS SubClass int64

MS Zoning object

Lot Frontage float64

Lot Area int64

Street object

Alley object

Lot Shape object

Land Contour object

Utilities object

Lot Config object

Land Slope object

Neighborhood object

Condition 1 object

Condition 2 object

Bldg Type object

House Style object

Overall Qual int64

Overall Cond int64

Year Built int64

Year Remod/Add int64

Roof Style object

Roof Matl object

Exterior 1st object

Exterior 2nd object

Mas Vnr Type object

Mas Vnr Area float64

Exter Qual object

Exter Cond object

Foundation object

Bsmt Qual object

Bsmt Cond object

Bsmt Exposure object

BsmtFin Type 1 object

BsmtFin SF 1 float64

BsmtFin Type 2 object

BsmtFin SF 2 float64

Bsmt Unf SF float64

Total Bsmt SF float64

Heating object

Heating QC object

Central Air object

Electrical object

1st Flr SF int64

2nd Flr SF int64

Low Qual Fin SF int64

Gr Liv Area int64

Bsmt Full Bath float64

Bsmt Half Bath float64

Full Bath int64

Half Bath int64

Bedroom AbvGr int64

Kitchen AbvGr int64

Kitchen Qual object

TotRms AbvGrd int64

Functional object

Fireplaces int64

Fireplace Qu object

Garage Type object

Garage Yr Blt float64

Garage Finish object

Garage Cars int64

Garage Area int64

Garage Qual object

Garage Cond object

Paved Drive object

Wood Deck SF int64

Open Porch SF int64

Enclosed Porch int64

3Ssn Porch int64

Screen Porch int64

Pool Area int64

Pool QC object

Fence object

Misc Feature object

Misc Val int64

Mo Sold int64

Yr Sold int64

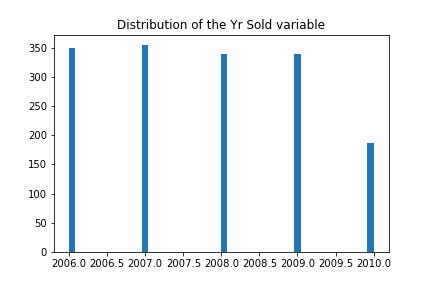
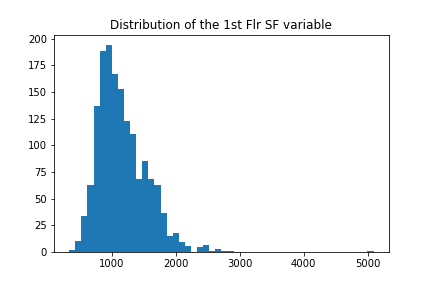
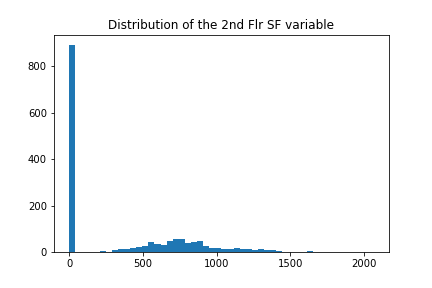
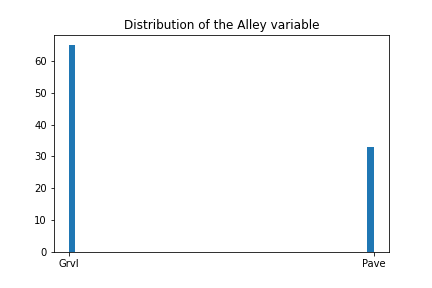
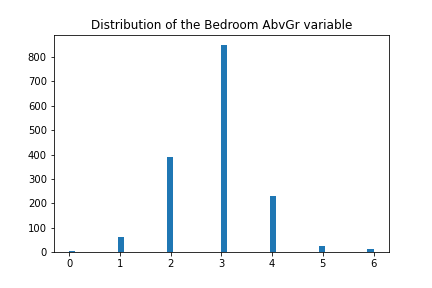
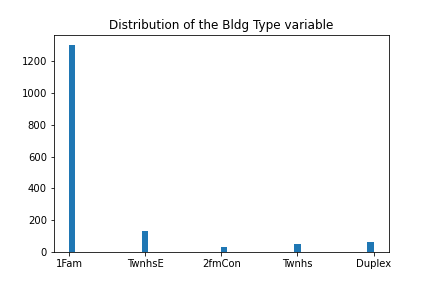
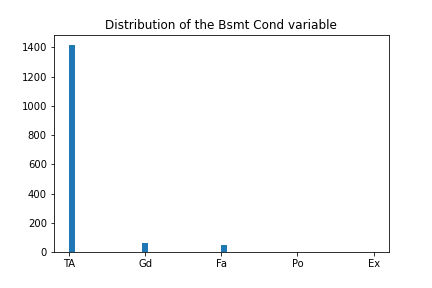
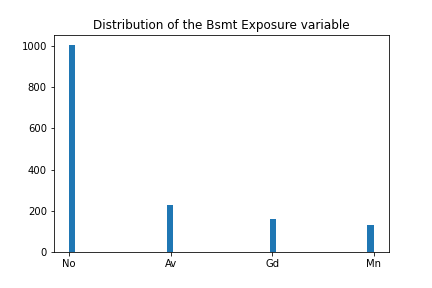
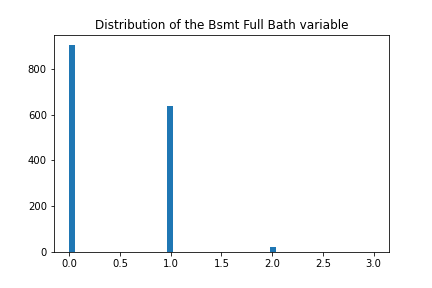
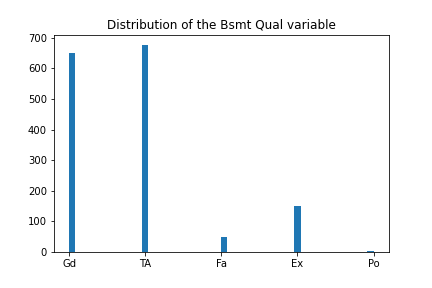
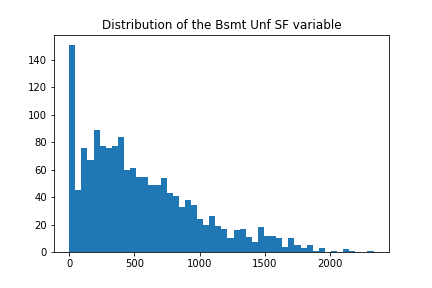
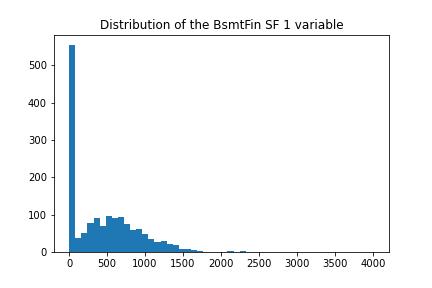
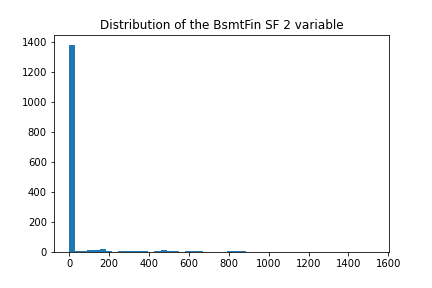
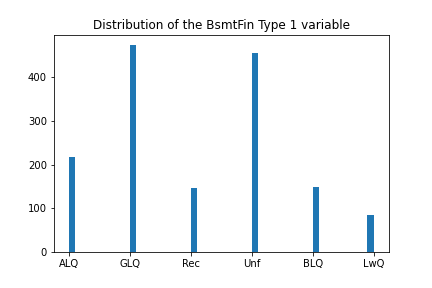
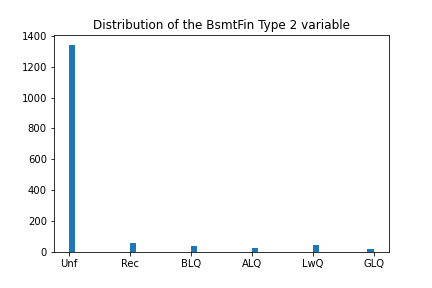
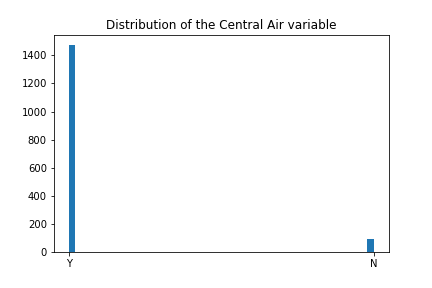
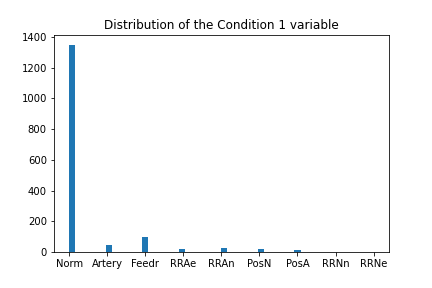
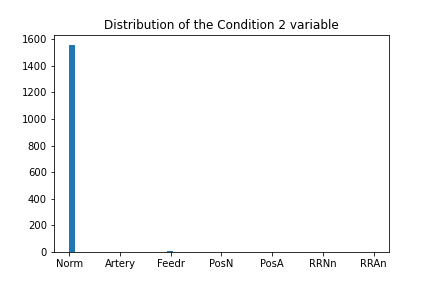
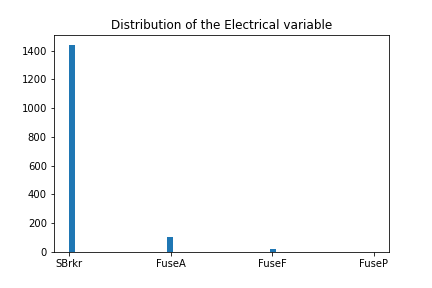
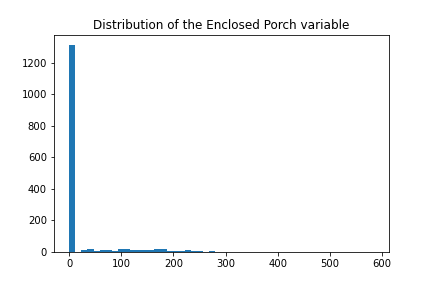
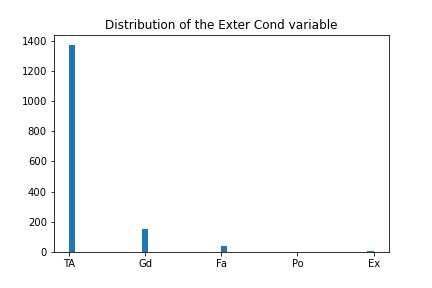
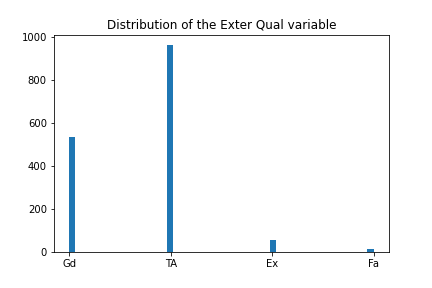
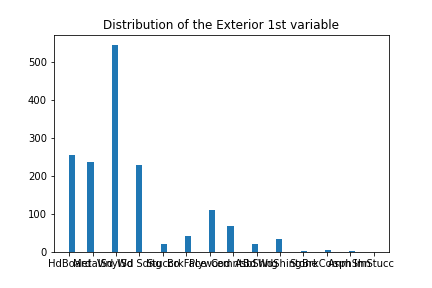
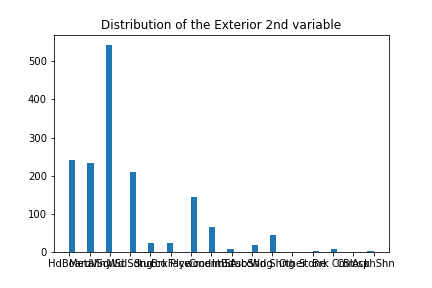
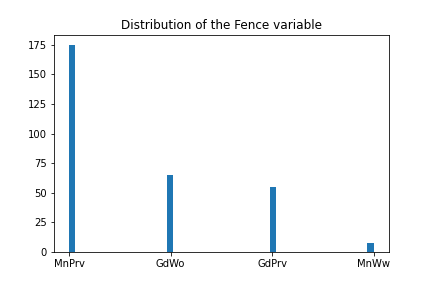
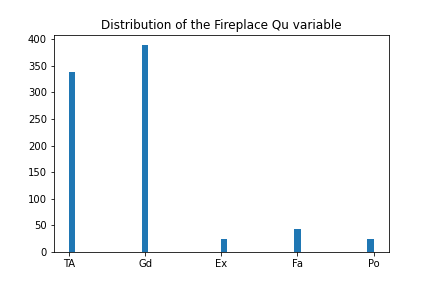
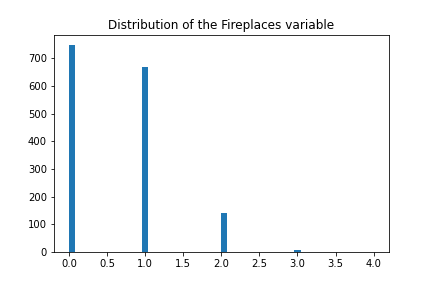
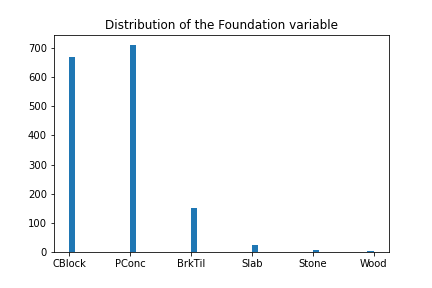
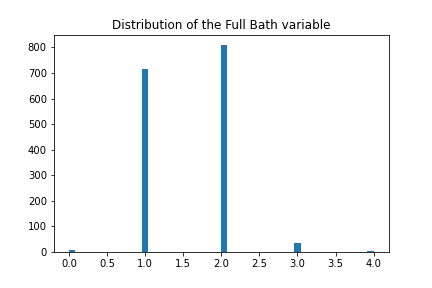
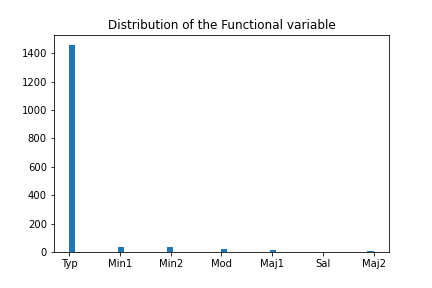
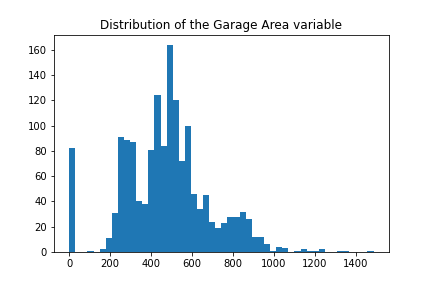
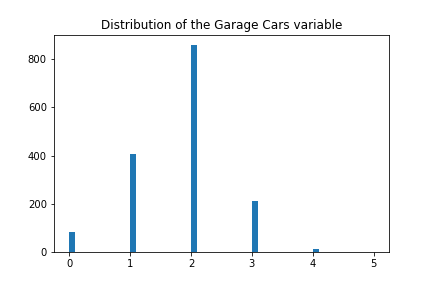
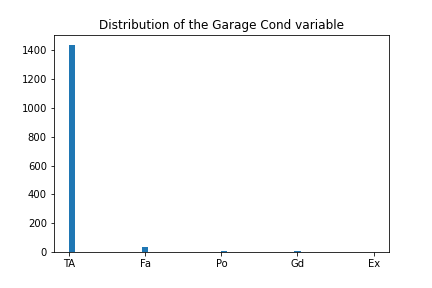
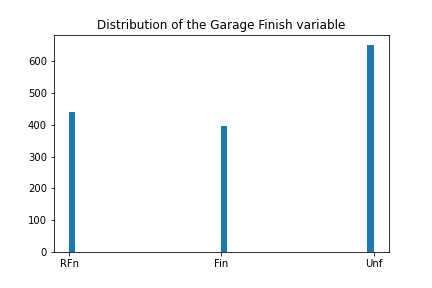
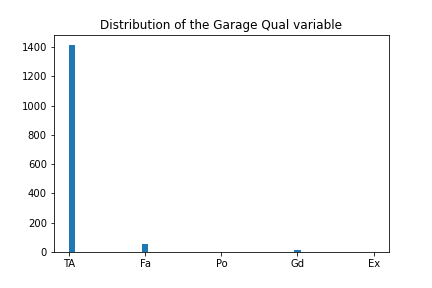
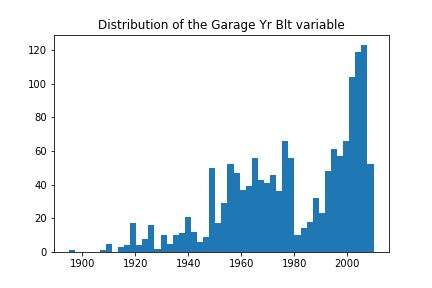
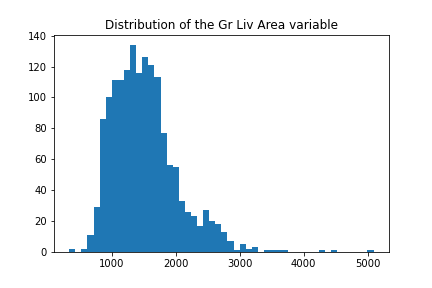
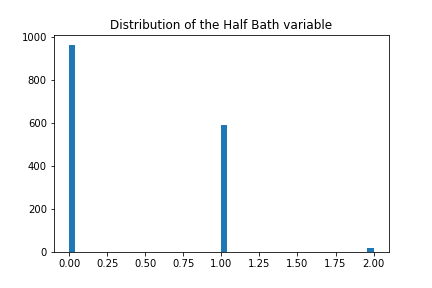
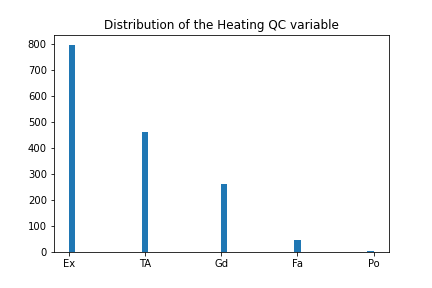
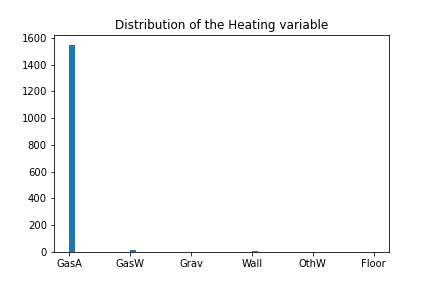
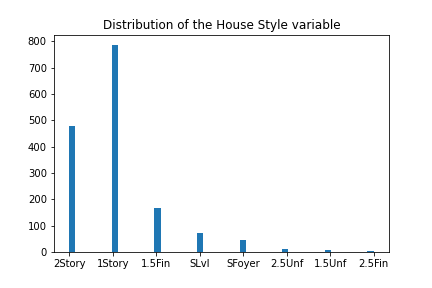
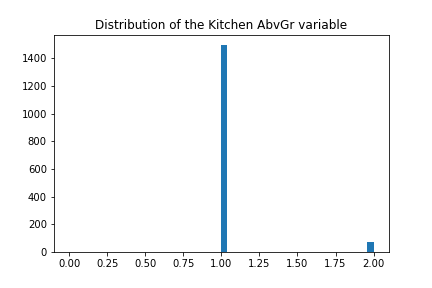
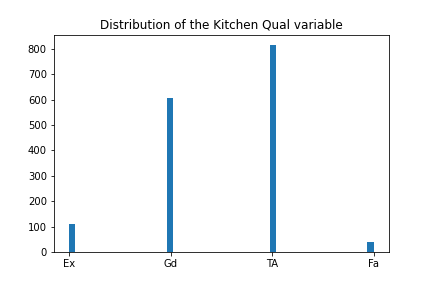
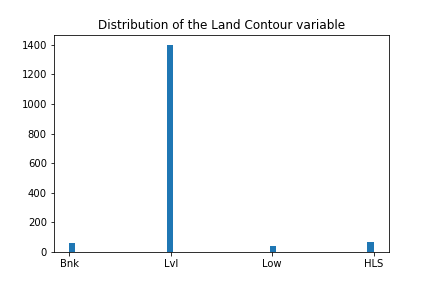
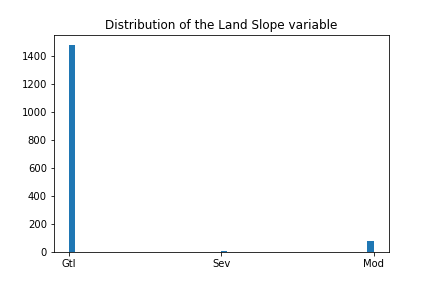
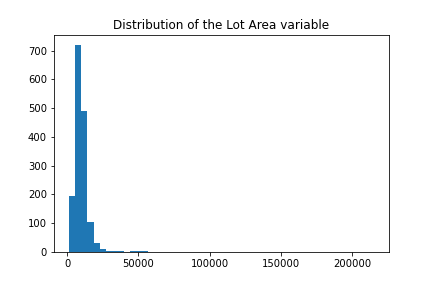
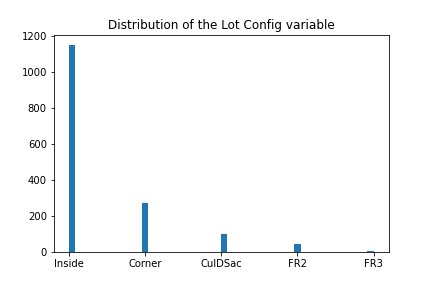
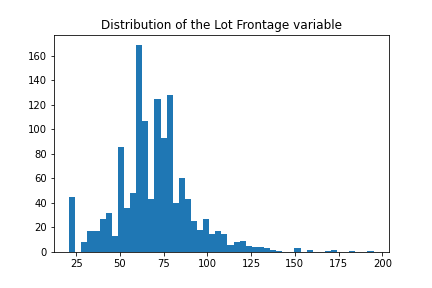
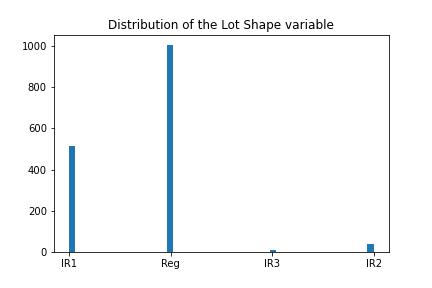
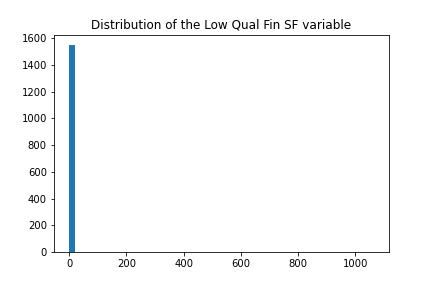
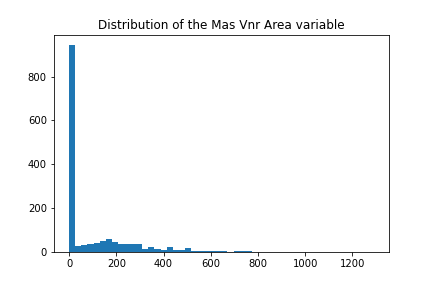
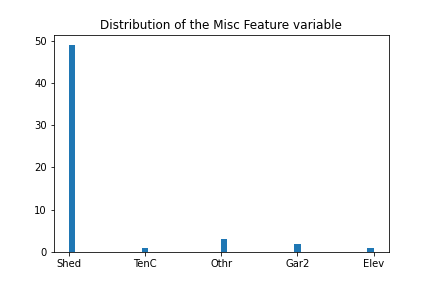
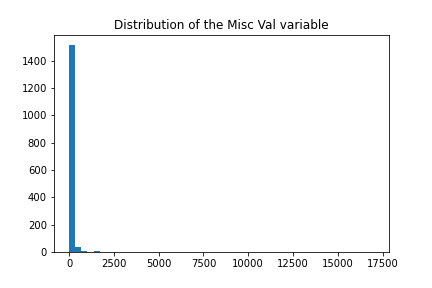
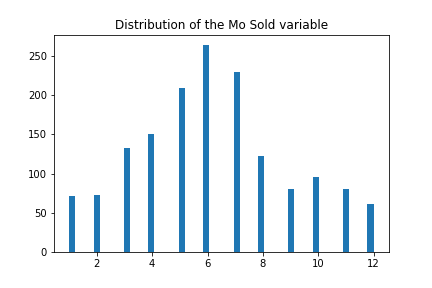
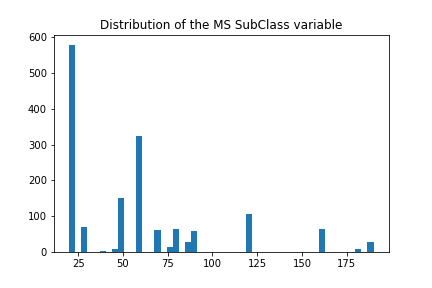
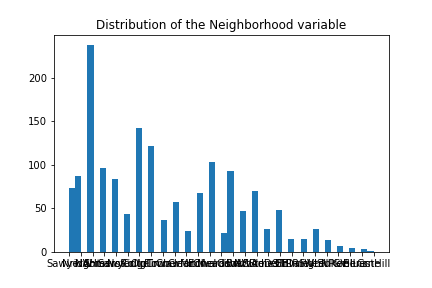
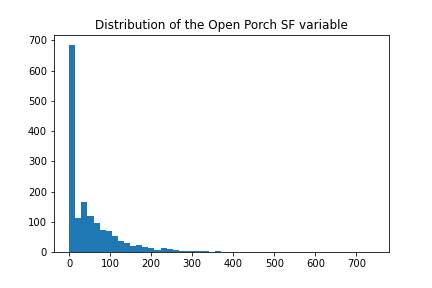
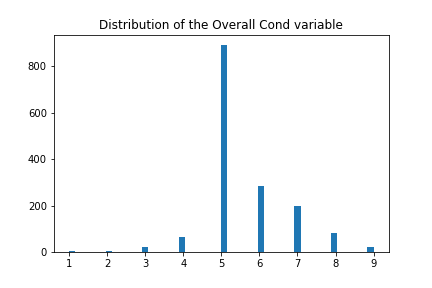
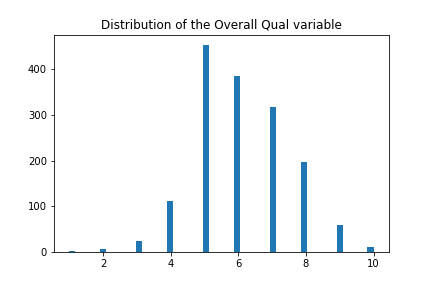
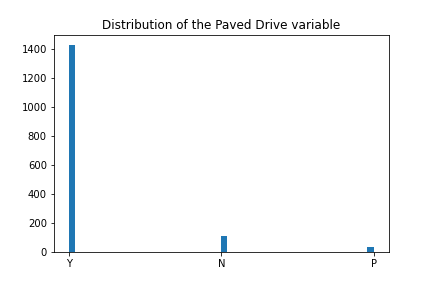
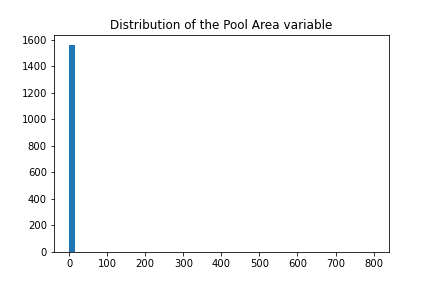
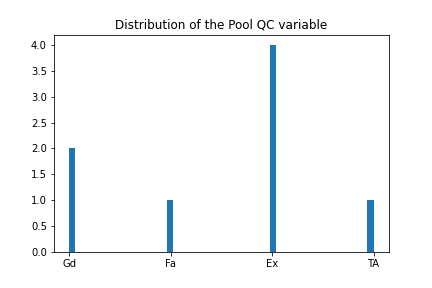
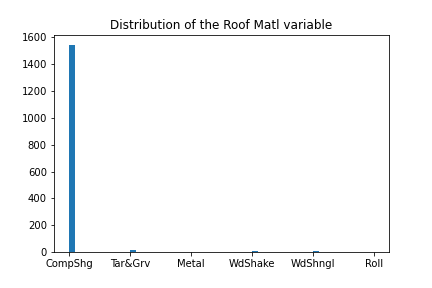
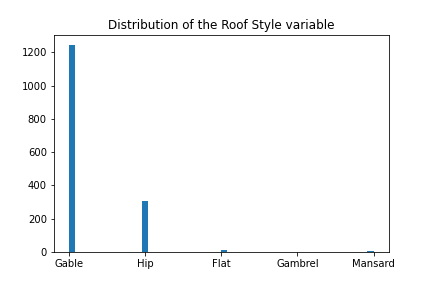
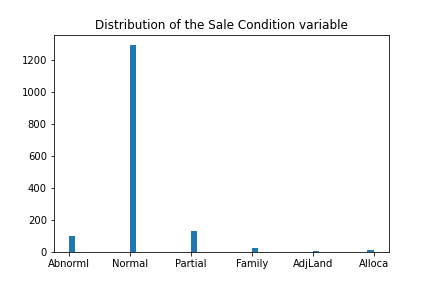
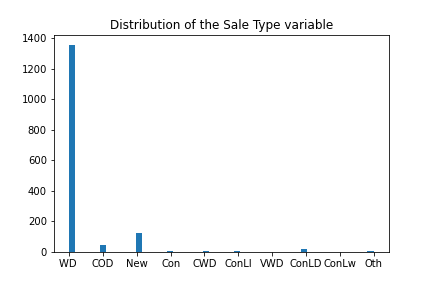
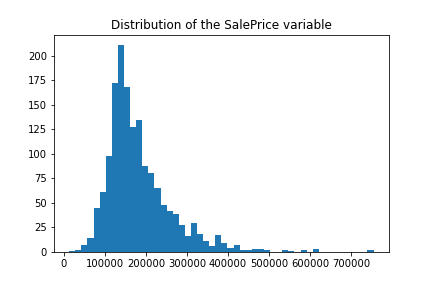
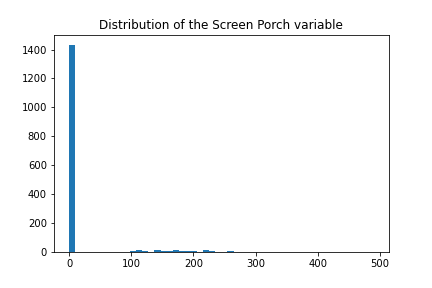
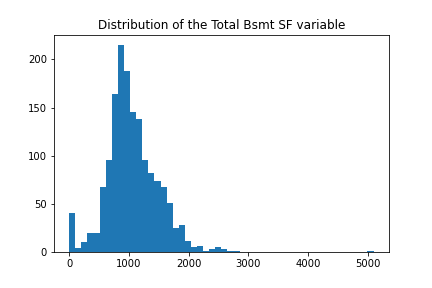
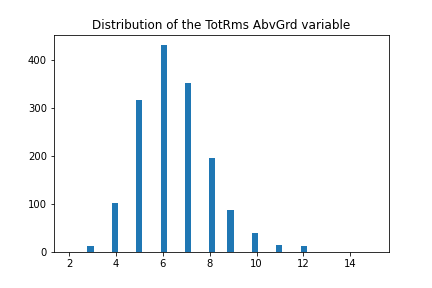
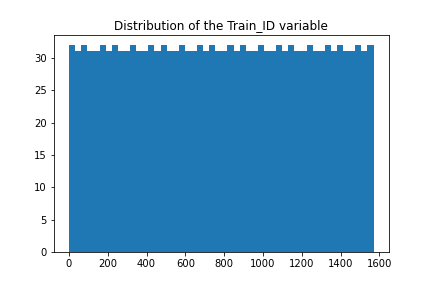
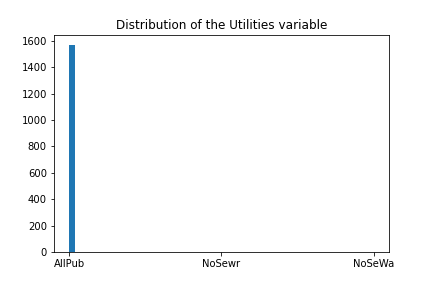
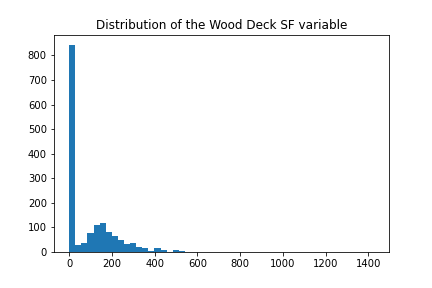
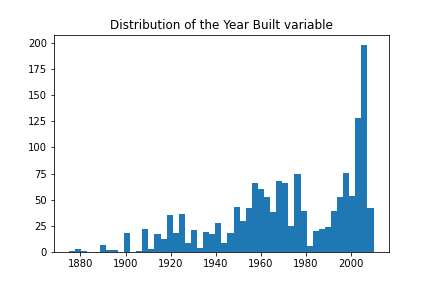
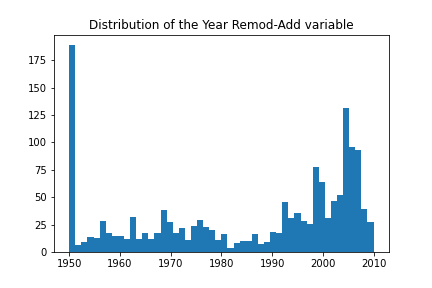
Sale Type object

Sale Condition object

SalePrice int64

dtype: object

## Appendix ii



## Appendix iii

## Appendix iv

## Appendix v

-------------------------------------------

---------------WITH OUTLIERS---------------

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best mlr model uses the features ['PER', 'DWS', 'Minutes', 'USG']

CV RMSE: 3.7293

Test RMSE: 4.1146

---------------WITH OUTLIERS---------------

-------------------------------------------

------OVERALL BEST MODEL BY TEST-RMSE------

-------------------------------------------

-------------------------------------------

Overall Best mlr model uses the features ['WS', 'Minutes']

CV RMSE: 4.0987

Test RMSE: 3.9841

…

-------------------------------------------

-------------WITHOUT OUTLIERS--------------

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best mlr model uses the features ['WS', 'Minutes', 'USG']

CV RMSE: 3.4782

Test RMSE: 4.0660

-------------WITHOUT OUTLIERS--------------

-------------------------------------------

------OVERALL BEST MODEL BY TEST-RMSE------

-------------------------------------------

-------------------------------------------

Overall Best mlr model uses the features ['WS', 'Minutes']

CV RMSE: 3.8000

Test RMSE: 3.9761

## Appendix vi

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best kNN model uses the features ['PER', 'DWS', 'Minutes', 'USG']

CV RMSE: 3.6202

Test RMSE: 4.0645

Number of neighbours: 11

-------------------------------------------

-------------------------------------------

------OVERALL BEST MODEL BY TEST-RMSE------

-------------------------------------------

-------------------------------------------

Overall Best kNN model uses the features ['PER', 'WS', 'Minutes']

CV RMSE: 3.9774

Test RMSE: 3.9741

Number of neighbours: 16

## Appendix vii

-------------------------------------------

------OVERALL BEST MODEL BY CV-RMSE--------

-------------------------------------------

-------------------------------------------

Overall Best svr model uses the features ['PER', 'DWS', 'Minutes', 'USG']

CV RMSE: 3.7372

Test RMSE: 4.0768

-------------------------------------------

------OVERALL BEST MODEL BY TEST-RMSE------

-------------------------------------------

-------------------------------------------

Overall Best svr model uses the features ['PER', 'WS', 'Minutes']

CV RMSE: 4.0052

Test RMSE: 4.0079

## Appendix viii

crim float64

zn float64

indus float64

chas int64

nox float64

rm float64

age float64

dis float64

rad int64

tax int64

ptratio float64

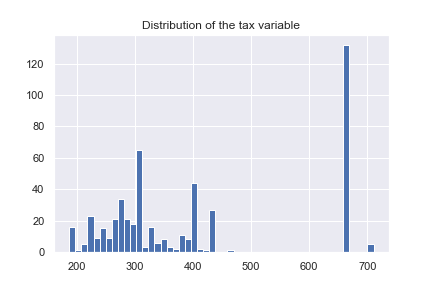
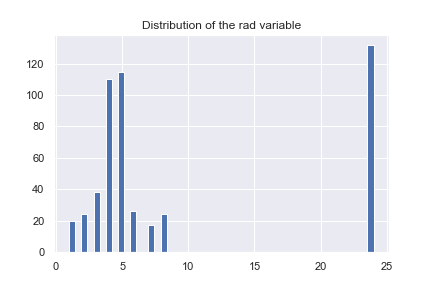
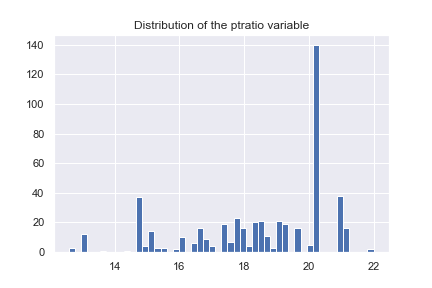
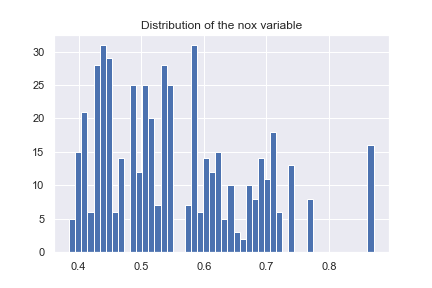
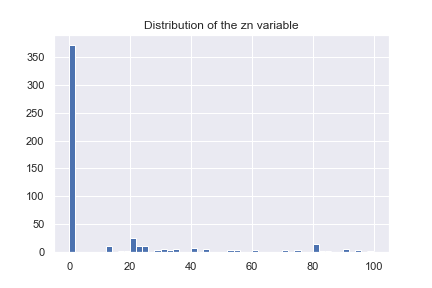
black float64

lstat float64

medv float64

dtype: object

## Appendix ix



## Appendix x

