Housing Sales Price Predictor

Data Mining - IE 7275

Group 1:
Unnati Ghodki
Anshita Aishwarya

The Team

UNNATI GHODKI

The "Tech-Savvy"

ANSHITA AISHWARYA

The "Jack-of-all-trades"

About the Project

The aim of the project is to estimate the housing sale prices with the highest accuracy by identifying the best model using data mining techniques.

OBJECTIVES

- To equip potential buyers, sellers, investors, insurance agencies, bankers, real estate marketers, etc. to understand the housing market conditions.
- To enable people in planning ahead for their "dream" houses.

Scope of the Project

What It Covers

- Data Collection & Processing
- Data Exploration & Visualization
- Model Exploration & Selection
- Implementation of Selected Model
- Performance Evaluation & Interpretation

PROJECT TIMELINE

01

BRAINSTORMING

Exploring vast dataset options and finalizing proposal

02

BUILDING BLOCKS

Planning the roadmap of project

03

EXECUTION

Deep dive into core Data Mining Concepts

04

ROADBLOCKS

Identifying the issues and troubleshooting

05

FINE-TUNING

Continuous development and integration

06

WRAP-UP

Presenting the analysis and reaching the finish line

Data Collection & Processing

Data Description

- Data is collected for residential homes in Ames, Iowa.
- Contains 79 explanatory variables as potential predictors (36 numerical and 43 categorical) and sales price as the target variable. The dataset contains 1460 records.
- Data Source: Kaggle

Variable Name	Description
Lot Area	Area of the lot (square feet)
Utilities	Type of utilities available - Electricity, Gas, Water, All
Building Type	Type of dwelling - Single family, Duplex, Townhouse
Bedroom	Number of bedrooms above basement level

Data Processing

1 **Date (Year) Columns:**

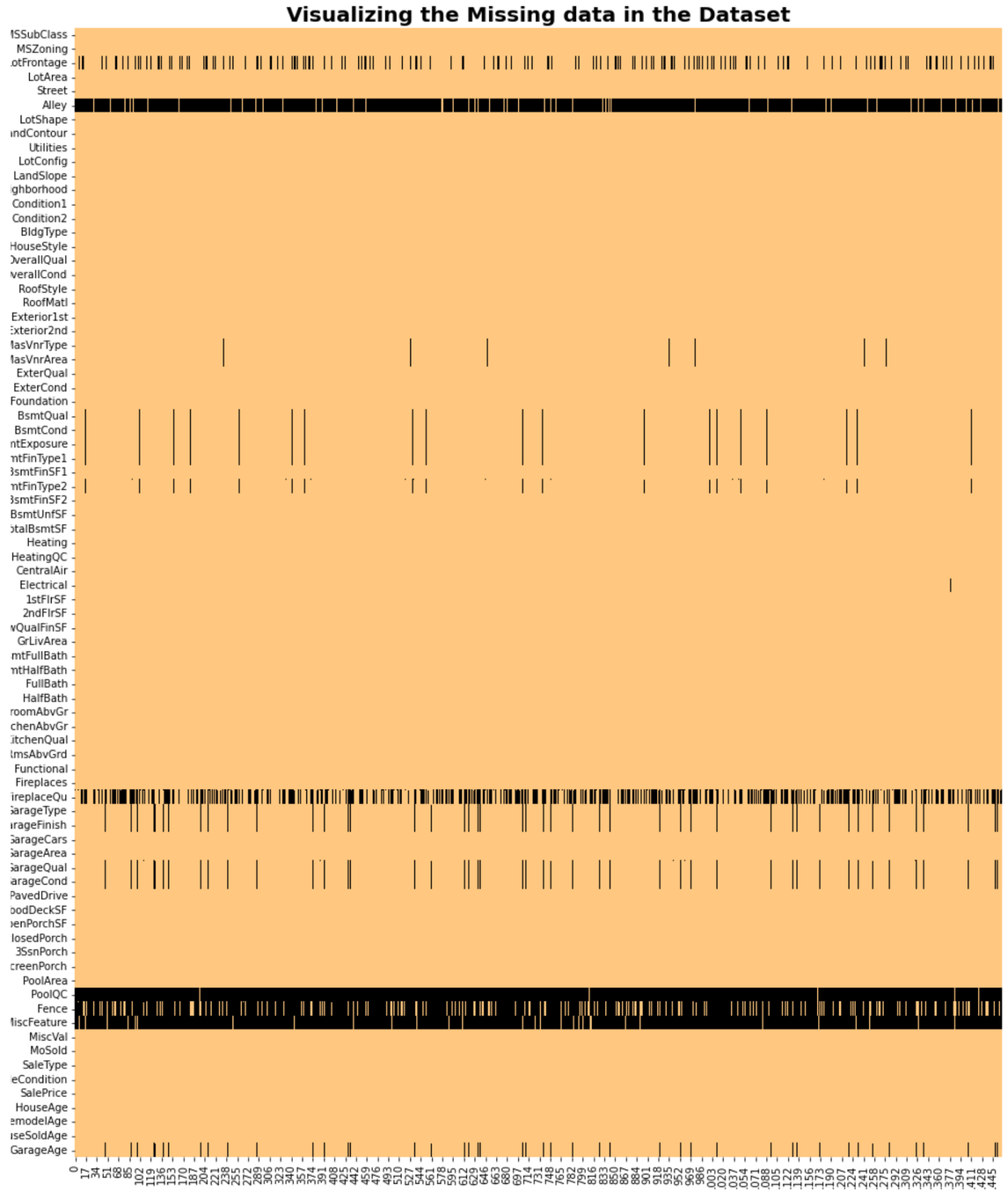
To have a more interpretable meaning, the year columns were converted to their respective ages up to today's year.

2 **Missing Values:**

- With 3 numerical columns with missing values, a threshold of 17% was used to drop the column. Other columns with missing values were imputed with 0.
- As for the categorical columns, the NA values represented the absence of that feature from the house. These values were imputed with 'None'.

3 **Dimensionality Reduction:**

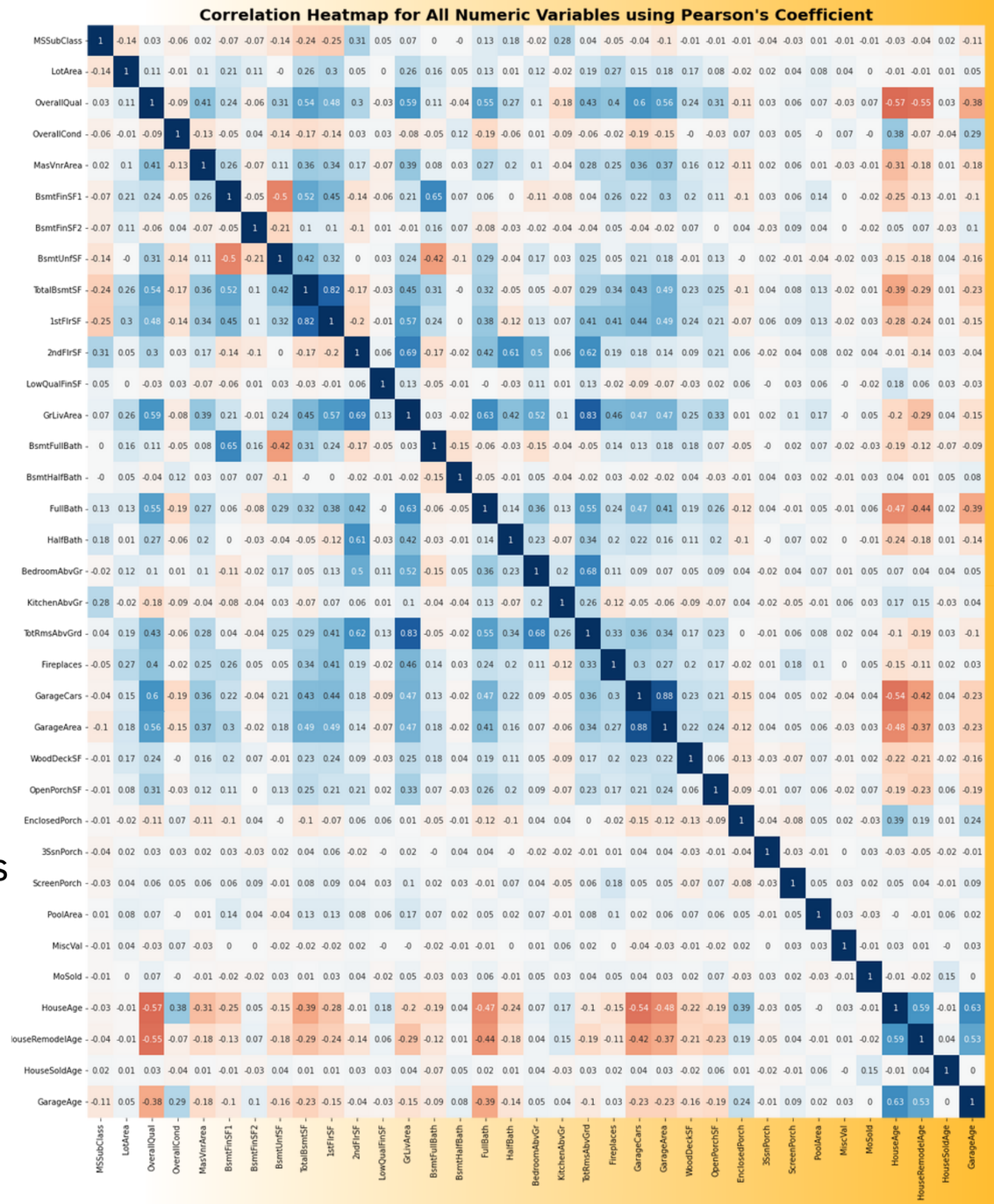
- *Pearson's correlation coefficient* is a bivariate correlation that measures the linear correlation between two sets of data, values ranging from -1 to 1.



- As part of dimension reduction, the correlation between every 2 numerical predictors has been calculated using Pearson’s coefficient and if a pair of variables are highly correlated (cutoff of 0.8), then one of the column-pair has been dropped.
- Similarly, ***Crammer's Rule*** is used to calculate the correlation between two categorical variables for selecting one category among the pairs in case of a higher correlation.

4 **Assessing Correlation between Predictors and Target variable:**

- The presence of uncorrelated predictors with the target variable increases the variance in the predicted output.
- A correlation coefficient analysis between all the predictors and the target variable has been performed and a threshold of 0.2 has been used to eliminate weakly correlated predictors.



- Using the dimension reduction techniques used above, there has been a successful reduction in predictors from 79 to 30 (12 categorical + 18 numerical), which contain significant information about the target variable.

5

Data Encoding:

- *Target Encoding* is a technique that encodes categorical predictors to numerical values since some models cannot handle categorical data.

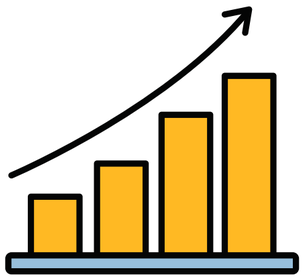
- This method results in only 1 dummy variable for a categorical column based on the mean value of the Target variable.

- This method has been chosen over *One-Hot Encoding* due to its benefit of not adding to the dimensionality of the dataset. Although “One-Hot Encoding” is an extremely easy technique to understand, it significantly increases the dimensionality of a dataset depending on the number of categories present in all the categorical columns.

Data Exploration & Visualization

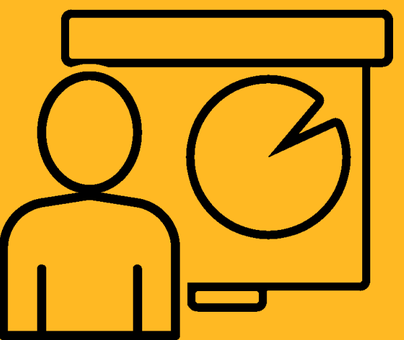
Data Exploration

- The different categories for each categorical feature were visualized using bar charts.
- It was observed that a lot of these predictors contained one major category and other less prevalent categories.
- 37 such columns were identified and the rare categories for each of these were combined to form a single category.
- A threshold of 20% of the most dominant category was fixed. That is, if any category has records 20% less than the most dominant category, then they were combined together to form a single "Other" category.



All the numerical predictors were visualized using histogram plots. Most of the predictors show a skewed distribution. The target variable Sale Price also has a right-skewed distribution.

Model Exploration & Selection



Data Preparation

1

Data Standardization:

The dataset contains values ranging from 1,300 to 215,245 square feet for the column *LotArea*.
On the other hand, it also contains single-digit values for *OverallQual* denoting the overall quality of the house.

Based on the fact that our data has a very large-scale difference, it has been standardized to bring the values on the same scale.

2

Data Splitting:

Before model exploration, data split has been performed to create partitioning into training and validation data.

75-25% ratio has been used for data partition.

Training Data		
X_train	1095	30
y_train	1095	1

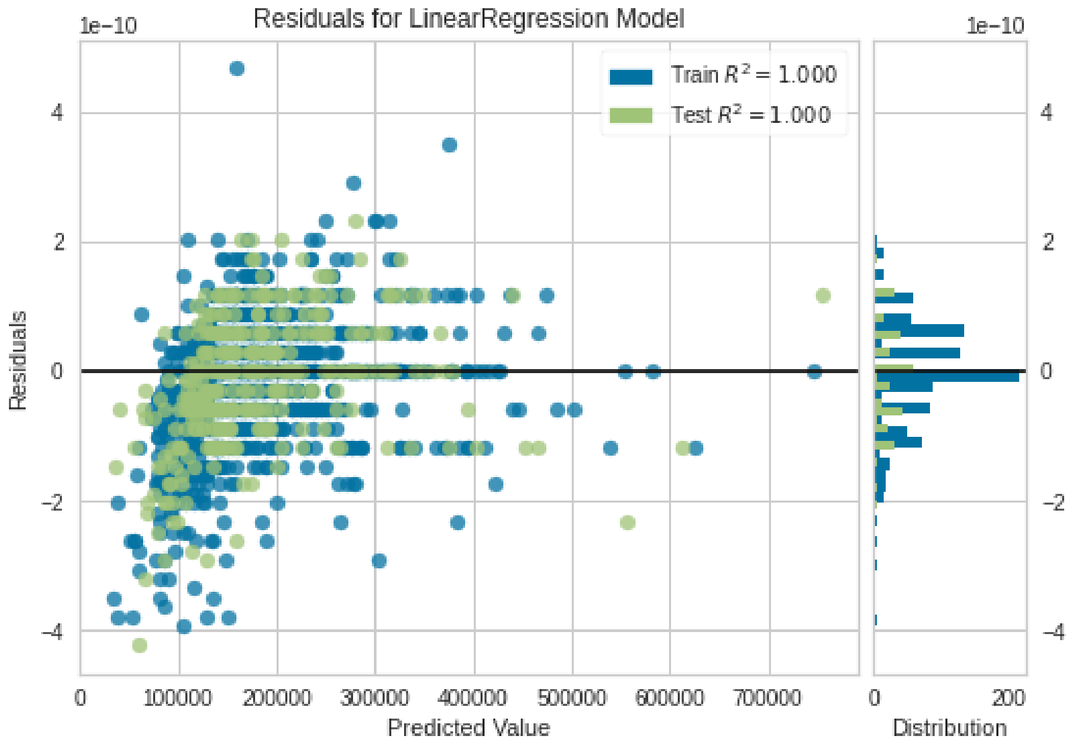
Validation Data		
X_test	365	30
y_test	365	1

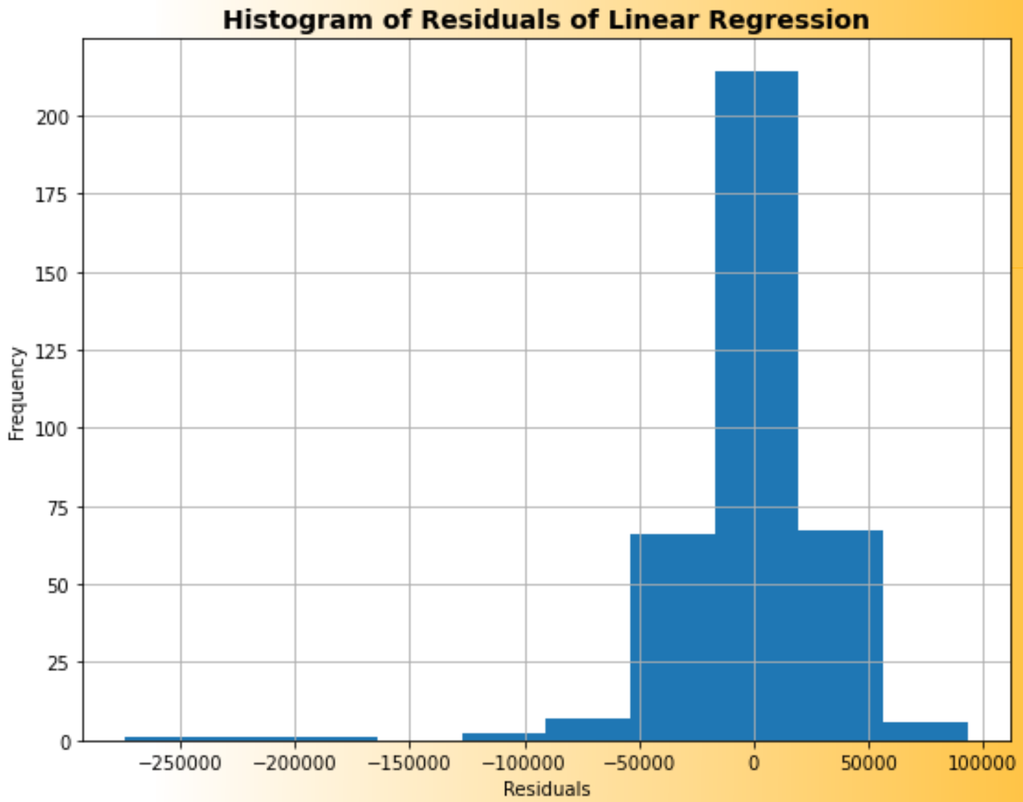
Model Exploration

1. Linear Regression
2. Regularized Linear Regression Model – Lasso
3. Regularized Linear Regression Model – Ridge
4. Regularized Linear Regression Model – Bayesian
5. Decision Trees
6. Random Forest

Linear Regression

Linear Regression is the machine learning model which fits the data assuming the linear relationship between predictors and target variable. Multiple linear regression has been implemented which uses *ordinary least squares* procedure to estimate the coefficients by minimizing the sum of squared errors between the predicted output and the actual output.

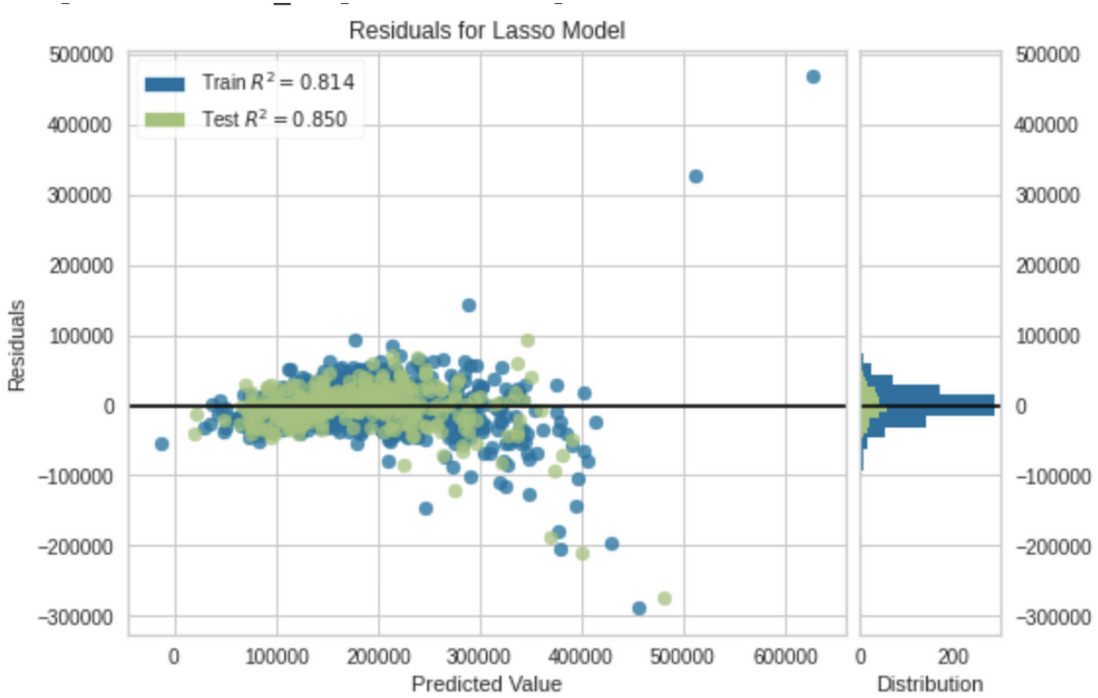


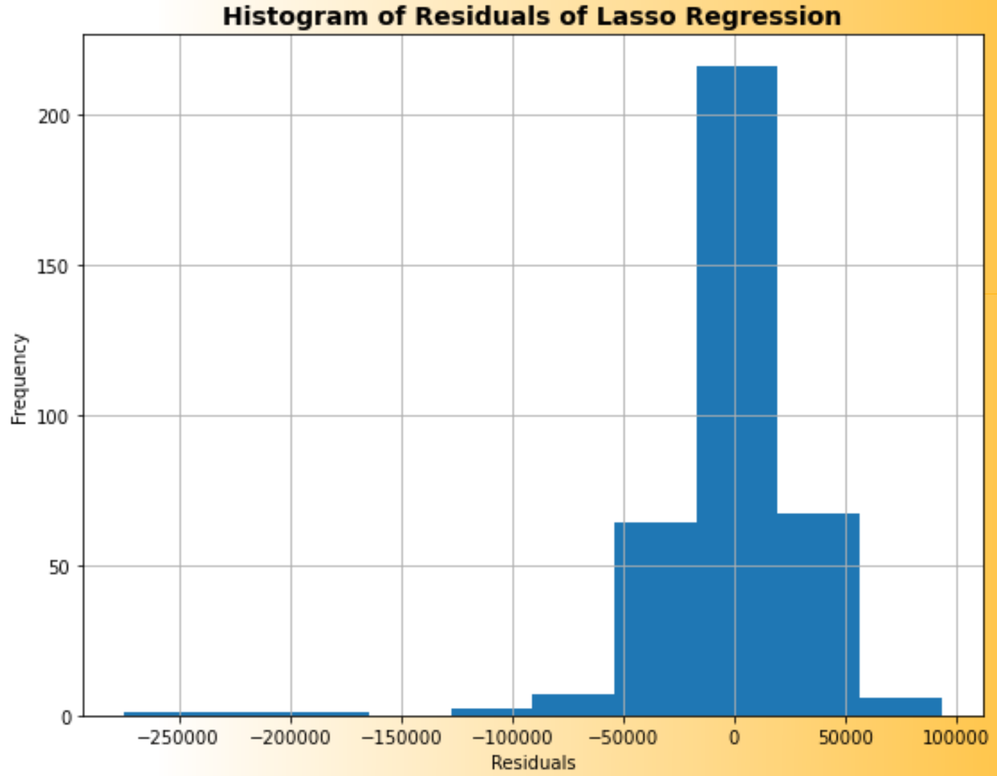


	MSE	RMSE	MAE	R2 Score
LinearRegression	1.049815e+09	32400.842	20517.168	0.814358

Regularized Linear Regression Model – Lasso

Lasso is a type of linear regression model which uses “shrinkage” wherein data points are shrunk towards any statistical measure. It uses L1 regularization and applies penalty based on the sum of absolute values of coefficients.

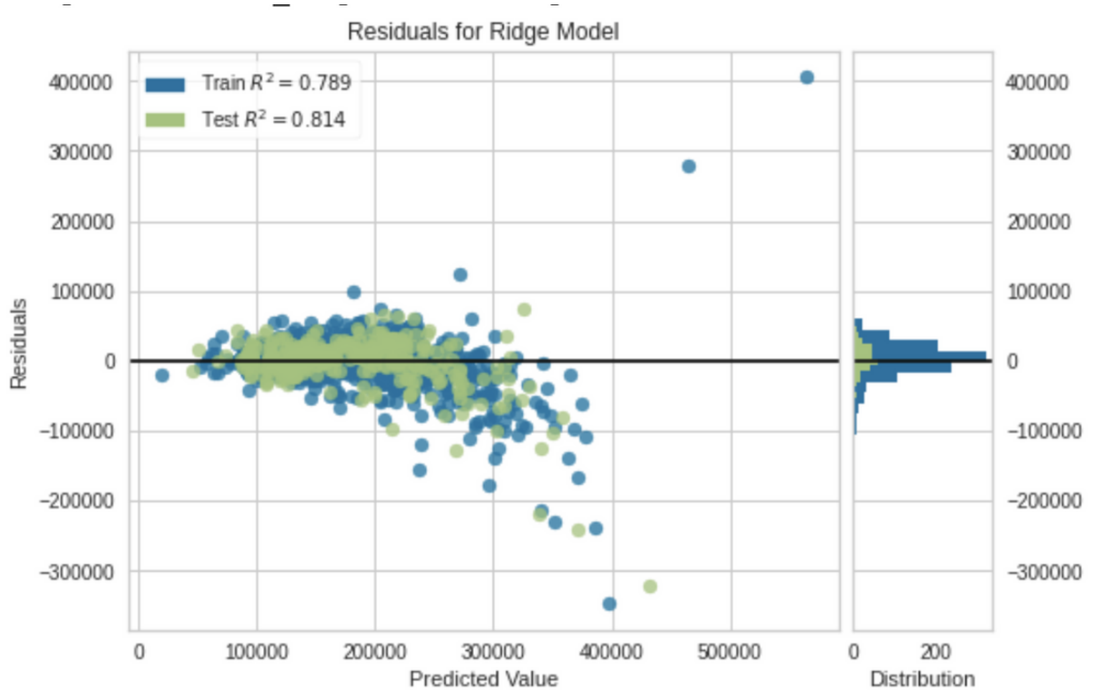


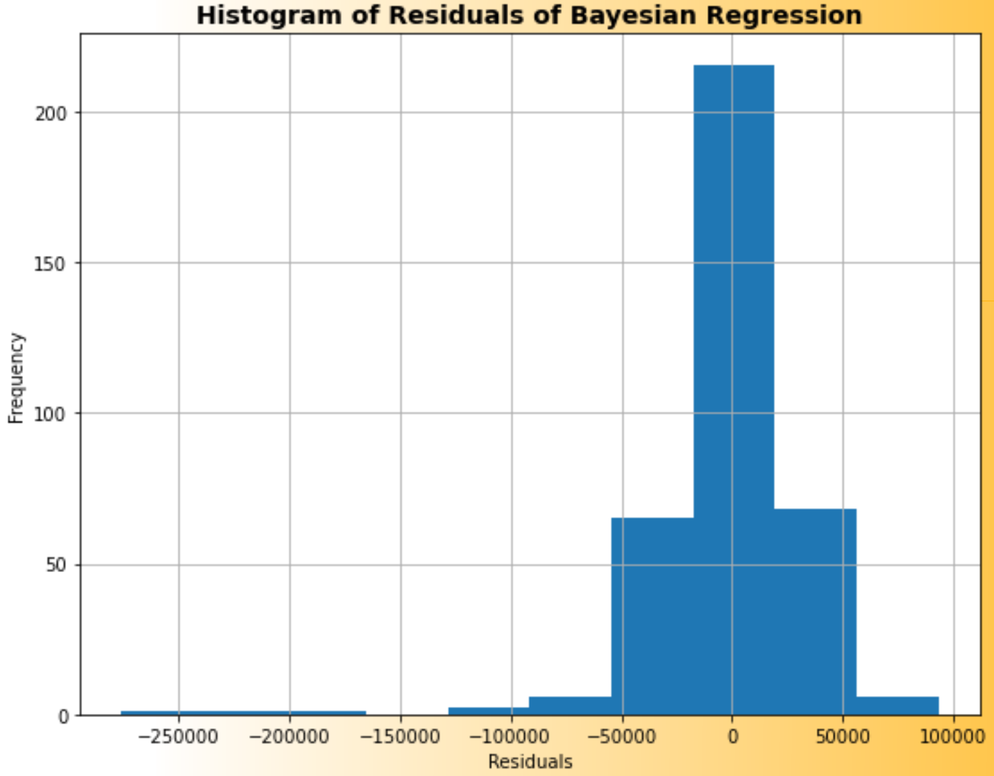


	MSE	RMSE	MAE	R2 Score
LassoRegression	1.052081e+09	32435.803	20545.669	0.849817

Regularized Linear Regression Model – Ridge

Ridge regression is a model tuning method that is analyzes data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

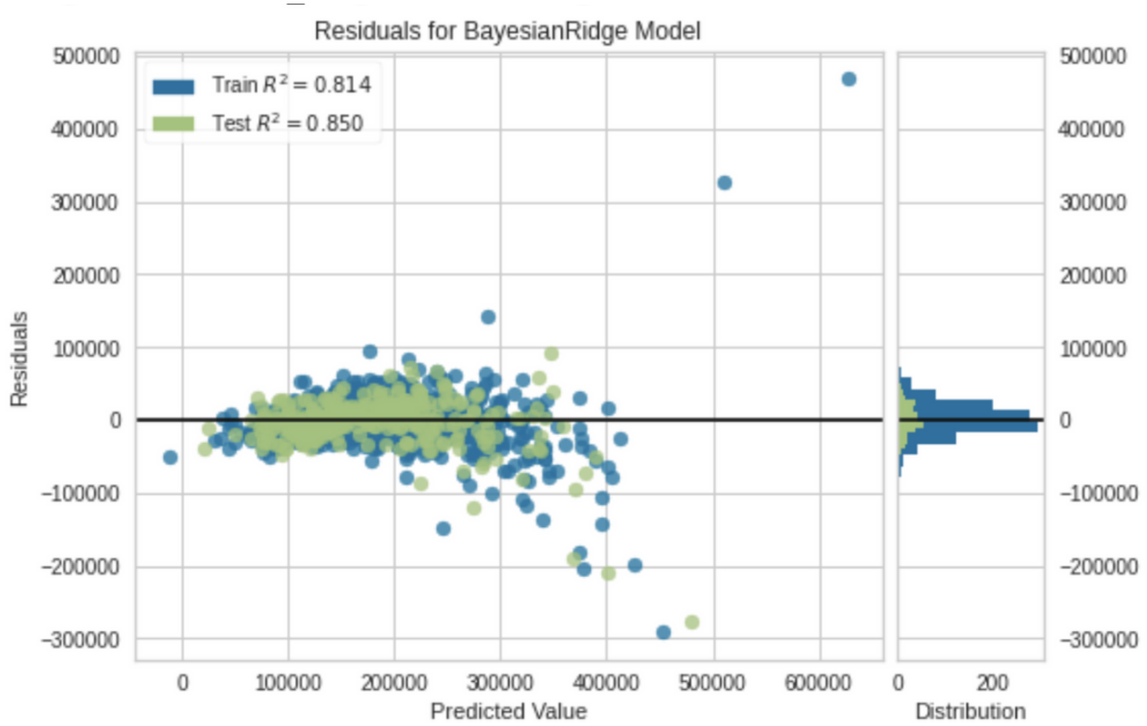


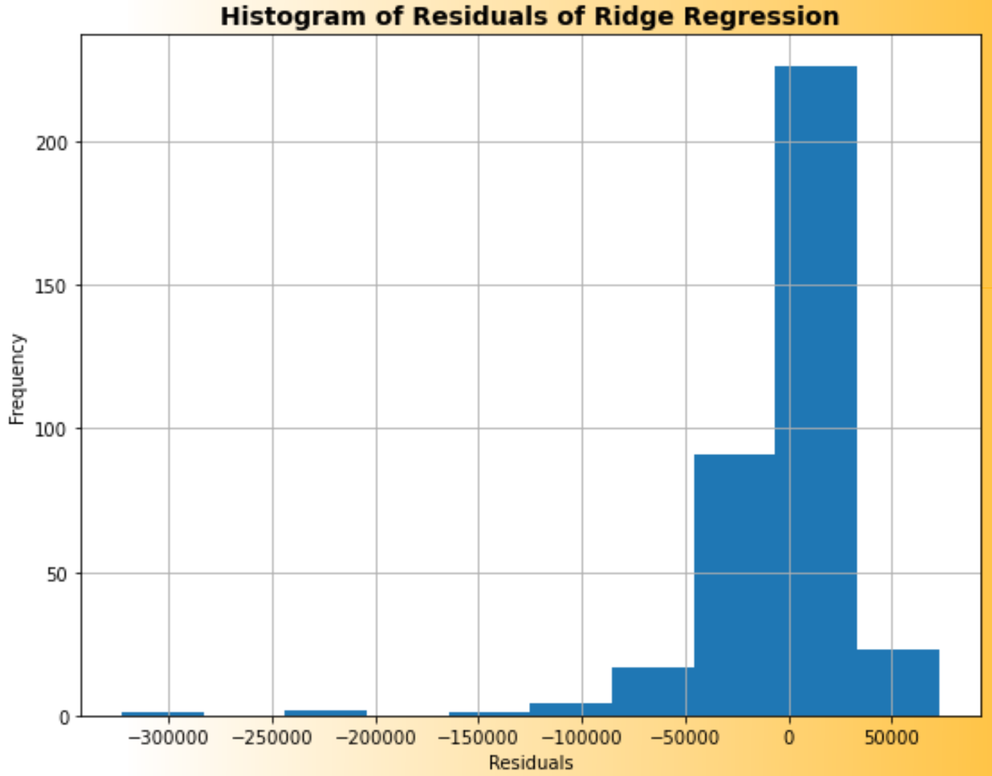


	MSE	RMSE	MAE	R2 Score
RidgeRegression	1.303697e+09	36106.741	21467.762	0.813899

Regularized Linear Regression Model – Bayesian

It uses Gaussian distribution to make the predictions instead of point estimates. We use probability distribution to estimate the target variable and the regression model is sampled from Normal distribution. The output variable is generated using Normal distribution metrics – means and variances.

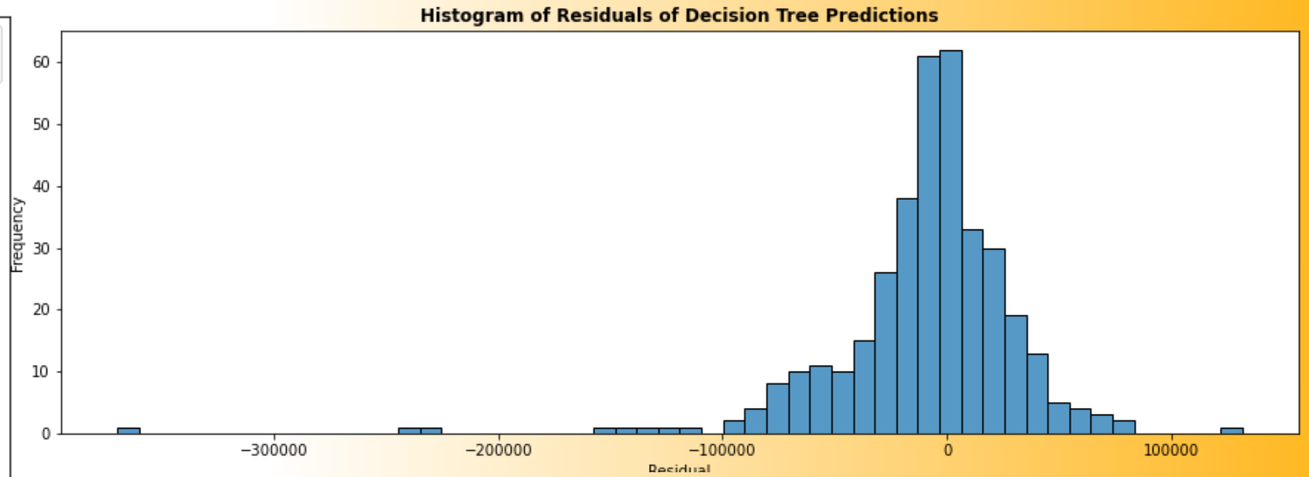




	MSE	RMSE	MAE	R2 Score
BayesianRegression	1.050558e+09	32412.312	20407.234	0.850034

Decision Trees

Decision tree is an algorithm that tries to strategically split a node into 2 or more sub-nodes in order to increase homogeneity of the resulting sub-nodes. Each node in the decision tree acts as a test case for some feature and based on the decision made at every node, the data is split into 2 or more sub-nodes. Decision trees end with leaf nodes that attempt to achieve maximum homogeneity without overfitting the data.

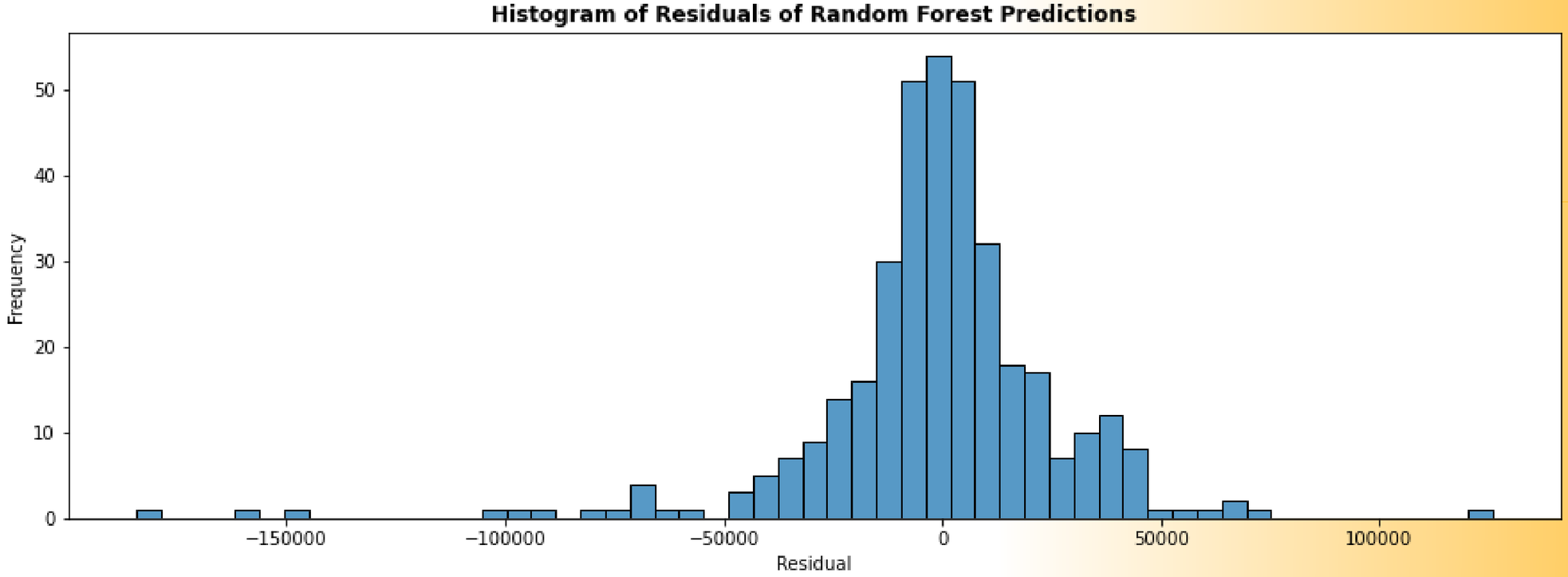


First of all, a grid search was performed to determine the optimal tree depth (number of edges from leaf node to the tree's root node). This has been done by calculating the mean squared error obtained by predicting the target values with varying tree depths. Ultimately, the model was presented with 5 as the optimal tree depth.

	MSE	RMSE	MAE	R2 Score
Decision_Tree	1.949306e+09	44150.944214	27318.750685	0.721739

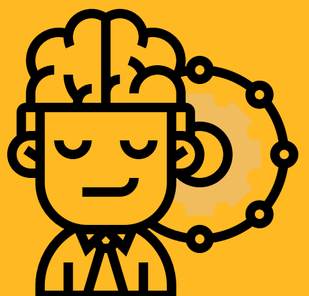
Random Forest

Random Forest is an ensemble of decision trees created using the bagging method (which states that a combination of learning models improves the overall method). Random forest adds randomness to the model while growing the trees, that is, while splitting a node, it searches for the best feature among a random subset of features. Therefore, in random forest, only a random subset of features are taken into consideration while fitting the model. This allows in obtaining better models as compared to decision trees.



	MSE	RMSE	MAE	R2 Score
Random_Forest	8.290347e+08	28792.962563	17979.587388	0.881656

Implementation of Selected Model



Model Selection

Models Explored	Mean Squared Error	Root Mean Squared Error	Mean Absolute Eroor	Coefficient of Determination (R2)
Linear Regression	1.182723e+09	34390.73	21610.87	0.791584
Lasso Regression	1.183878e+09	34407.53	21603.39	0.831003
Ridge Regression	1.508248e+09	38836.17	22705.81	0.784699
Bayesian Regression	1.184527e+09	34416.95	21512.87	0.83091
Decision Tree	1949306000	44150.94	27318.75	0.721739
Random Forest	**807492500**	**28416.41**	**18047.66**	**0.884731**

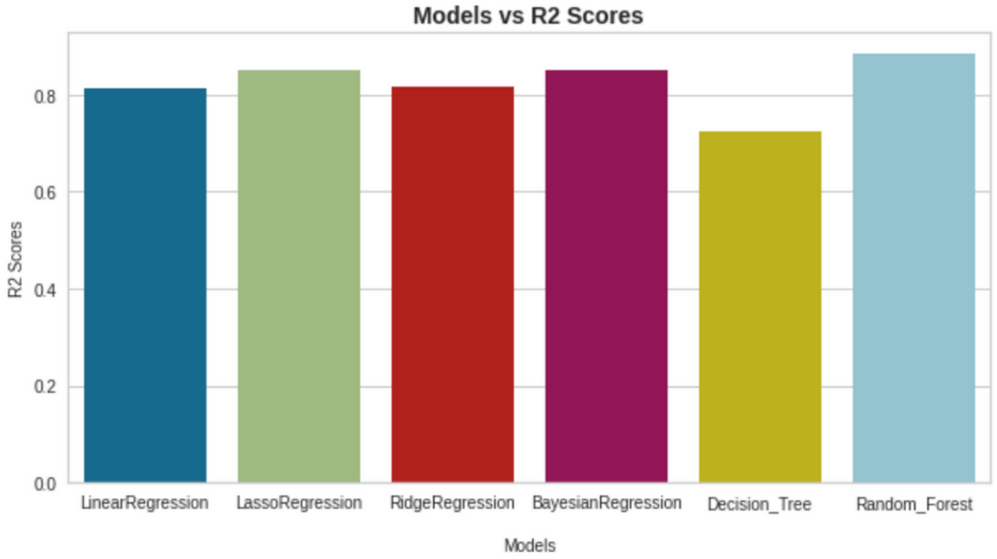
Based on the different models explored and their respective performance metrics, **Random Forest** has been performing in the best way by reducing all kind of errors and hence, is identified as the best model for this dataset.

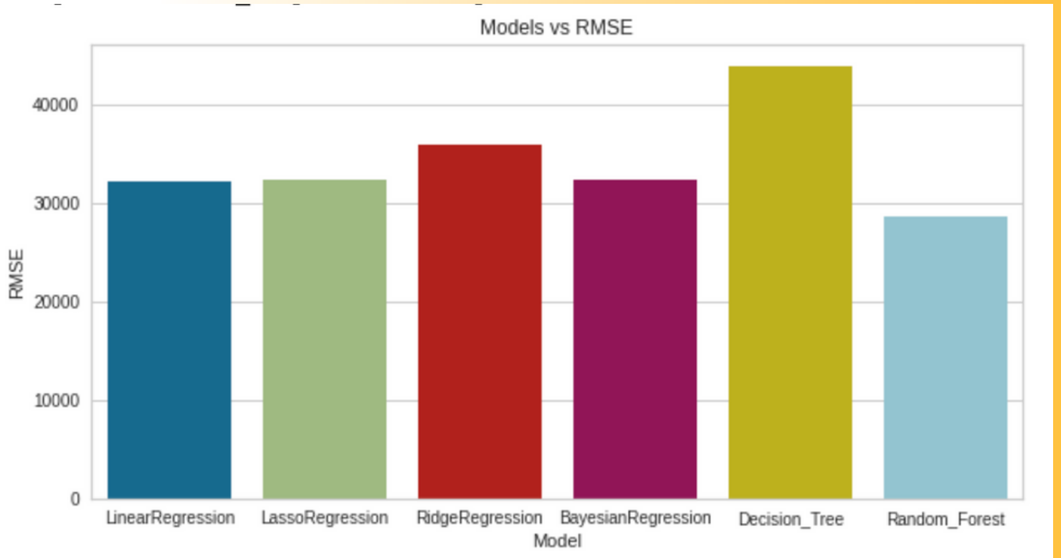
Advantages of using this model:

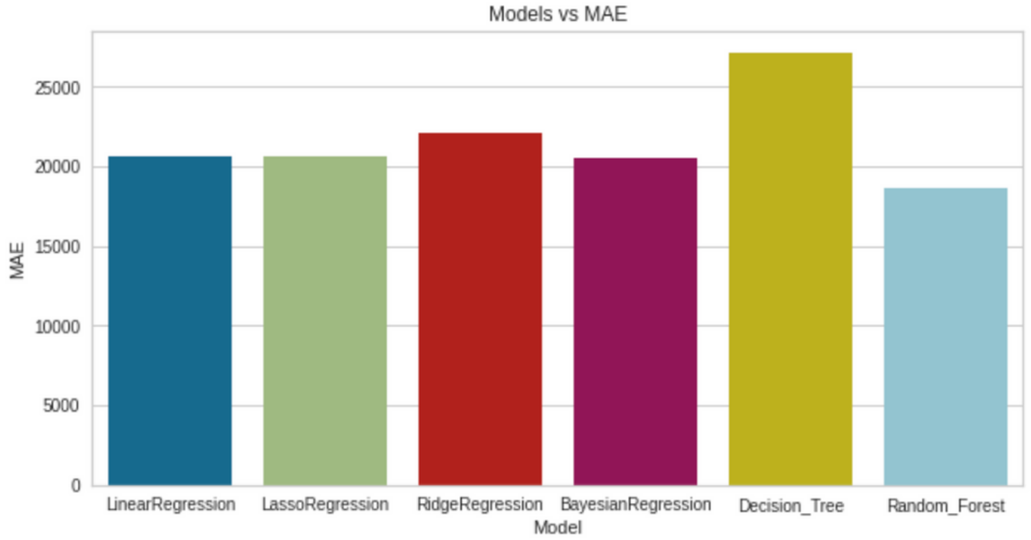
- Since Random Forest is based on the bagging algorithm, it creates many trees by choosing random features from the dataset and thus is effective in improving the accuracy and is resilient against overfitting.
- Random Forest works well with both categorical and numerical variables.
- It has the ability to handle the non-linear parameters efficiently.
- Random Forest can automatically handle missing values.
- It is a robust model in dealing with outliers.
- The algorithm is stable and this model works really well even when a new data point is introduced.
- Random Forest is comparatively less impacted by noise.

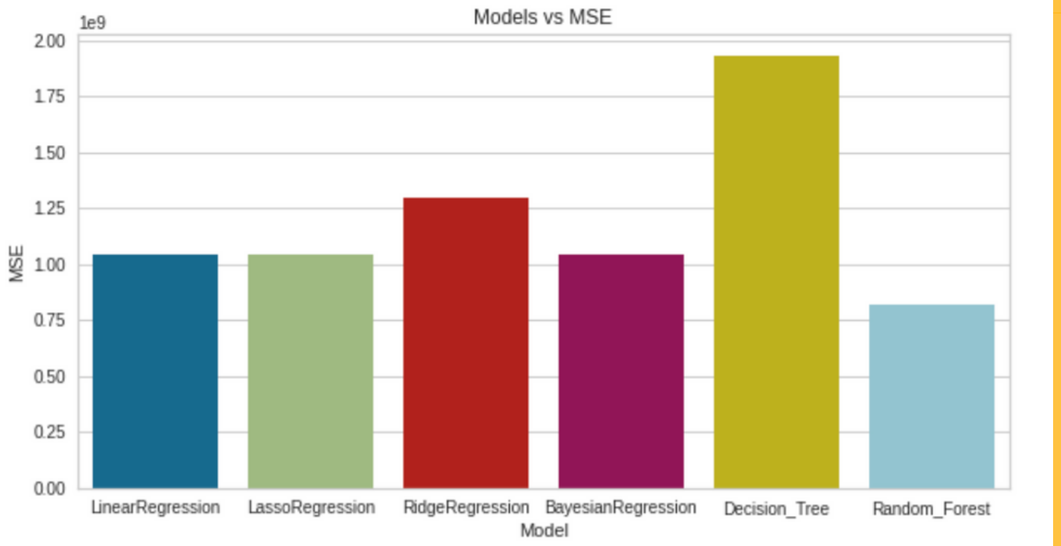
Performance Evaluation & Interpretation

Performance of Different Models





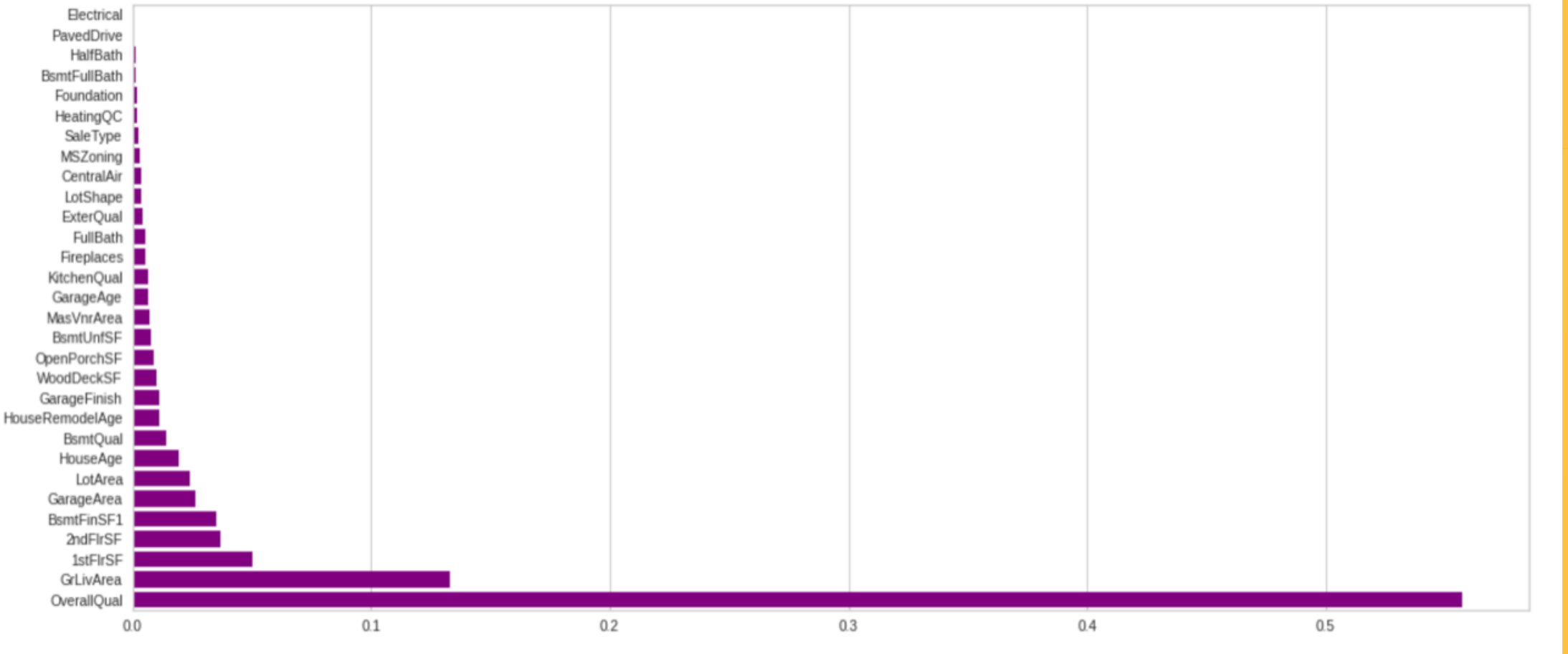




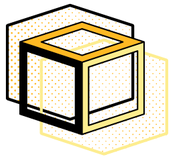
Interpretation

Best Features - Random Forest Classifier

1. Overall Quality of the House
2. Ground Living Area
3. 1st Floor Square Feet
4. 2nd Floor Square Feet



Challenges Faced During Execution



Dimensionality

The vast dimensionality of the dataset, which led us to explore various dimension reduction and feature selection methods.

One Hot Encoding

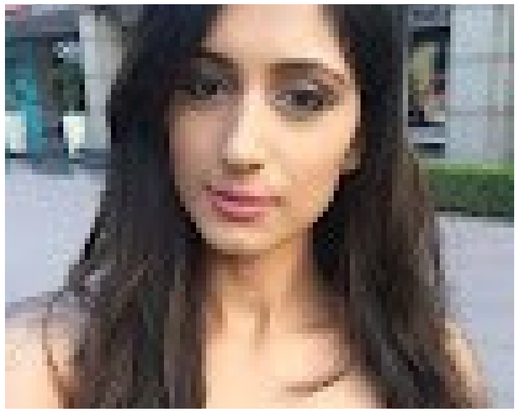
The one-hot method increased the the dimensionality significantly which led to overfitting the data. Thus, target encoding was performed.

High Computation Time

Since the dataset was very large, various code execution (visualizing features, correlation analysis etc.) were taking longer time to execute.

Project Resources - People

Professor

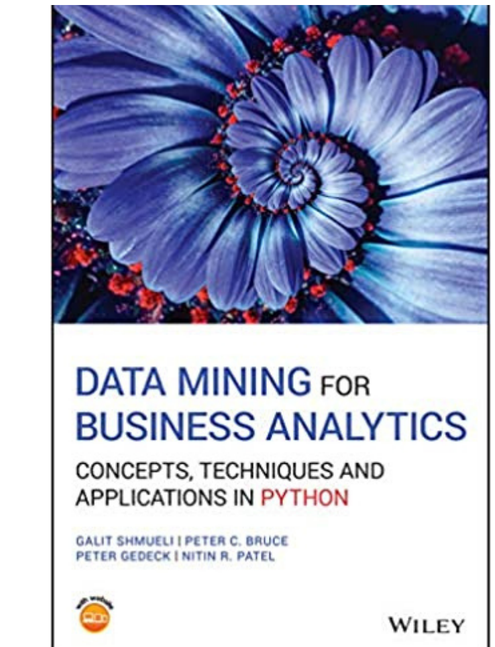


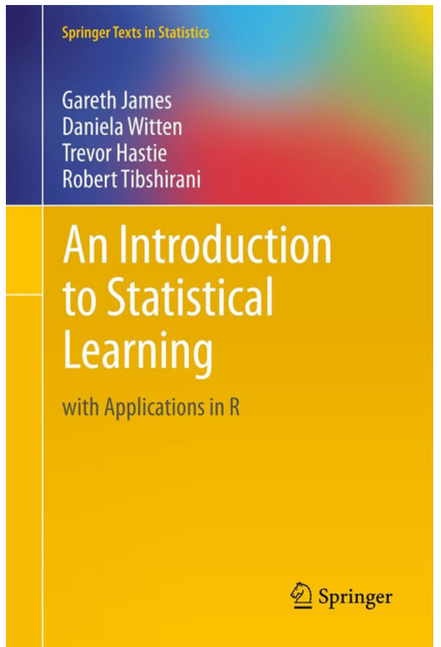
TA

Resource Page

Online Resources :

Textbooks :





Thank you!