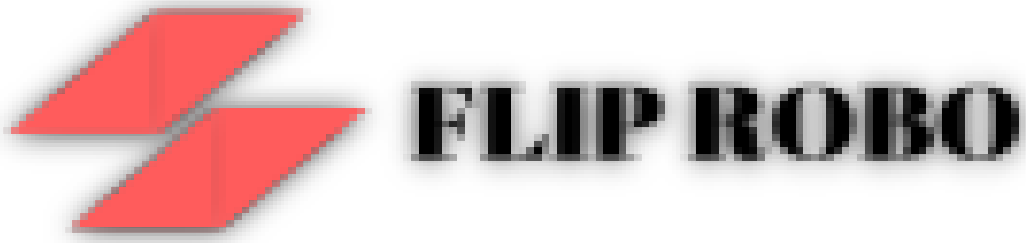


House Price Prediction

Submitted by:

Utsav Rastogi

ACKNOWLEDGMENT

This project was made possible with the help of my supervisor Swati Mahaseth. Under her guidance, I was able to complete the project on time and ample learning possibilities this project provided me with.

This project help me attain knowledge of managing datasets with large number of features and also redefine the usual process I've been following for each Machine Learning Project.

Data analysis and preprocessing was done keeping in mind how to make the code run efficiently and quickly with utilizing minimal memory. I had to refer to various articles on Medium, Analytics Vidhya, Kaggle and stack exchange to look for potential debugging solutions and visualization plots.

INTRODUCTION

Business Problem Framing

This project tries to shed light on an efficient way to predict the price of houses in Australia. This project meant to help understanding the various factors at play when predicting a price of house in Australia.
The predictive model can help the company optimally price their properties to attract customers, and enhance their marketing strategy to outcompete other players from the market.

Conceptual Background of the Domain Problem

Real estate comprises of various sectors such as commercial, residential and industrial. Pricing such properties cannot be done on just the human instinct, it needs to be done statistically and provide an optimal price for both the customer and the company.

Review of Literature

This dataset was taken from an online competition on Kaggle. This data was further analyzed and cleaned with standard pre-processing steps which will be subsequently used to create a predictive model. The model predicts the price of the house using several independent features. The model is based on regression algorithms which are usually used to predict a continuous target variable.

The model with least difference in r2 score and cross validation score is finalized and saved in a pickle file. This model can be further integrated by webapps to check price of any house in Australia.

Motivation for the Problem Undertaken

The objective behind this project was to provide the client with an overview of how house prices can be marked efficiently with a statistical model. The growth of Real Estate sector is well complemented by the growth of the corporate environment and the demand for office space as well as urban and semi-urban accommodations.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

Statistical analysis	Zscore was used to determine the outliers in the data and Correlation helped understand the relation between features and target. Skewness was also used to check the distribution and reduce it in the features which weren't gaussian.
Visual analysis	Libraries used: Matplotlib and Seaborn
Algorithms	Different Regression algorithms such as XGB Regressor & Lasso Regression were used which are all provided by sklearn library.

Data Sources and their formats

1. Data is sourced from:

   Data has been made available by a US-based housing company on Kaggle

2. Data is available in CSV format
3. Data Features:
   - Data consists of both numerical and categorical columns.

MSSubClass	Identifies the type of dwelling involved in the sale
MSZoning	Identifies the general zoning classification of the sale
LotFrontage	Linear feet of street connected to property
LotArea	Lot size in square feet
Street	Type of road access to property
Alley	Type of alley access to property
LotShape	General shape of property
LandContour	Flatness of the property

Utilities	Type of utilities available
LotConfig	Lot configuration
LandSlope	Slope of property
Neighborhood	Physical locations within Ames city limits
Condition1	Proximity to various conditions
Condition2	Proximity to various conditions (if more than one is present)
BldgType	Type of dwelling
HouseStyle	Style of dwelling
	Rates the overall material and finish of the house
OverallCond	Rates the overall condition of the house
YearBuilt	Original construction date
OverallQual YearRemodAdd	Remodel date (same as construction date if no remodeling or additions)
RoofStyle	Type of roof
RoofMatl	Roof material
Exterior1st	Exterior covering on house
Exterior2nd	Exterior covering on house (if more than one material)
MasVnrType	Masonry veneer type
MasVnrArea	Masonry veneer area in square feet
ExterQual	Evaluates the quality of the material on the exterior
ExterCond	Evaluates the present condition of the material on the exterior
Foundation	Type of foundation
BsmtQual	Evaluates the height of the basement
BsmtCond	Evaluates the general condition of the basement
BsmtExposure	Refers to walkout or garden level walls
BsmtFinType1	Rating of basement finished area
BsmtFinSF1	Type 1 finished square feet
BsmtFinType2	Rating of basement finished area (if multiple types)

BsmtFinSF2	Type 2 finished square feet
BsmtUnfSF	Unfinished square feet of basement area
TotalBsmtSF	Total square feet of basement area
Heating	Type of heating
HeatingQC	Heating quality and condition
CentralAir	Central air conditioning
Electrical	Electrical system
1stFlrSF	First Floor square feet
2ndFlrSF	Second floor square feet
LowQualFinSF	Low quality finished square feet (all floors)
GrLivArea	Above grade (ground) living area square feet
BsmtFullBath	Basement full bathrooms
BsmtHalfBath	Basement half bathrooms
FullBath	Full bathrooms above grade
HalfBath	Half baths above grade
Bedroom	Bedrooms above grade (does NOT include basement bedrooms)
Kitchen	Kitchens above grade
KitchenQual	Kitchen quality
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
Functional	Home functionality (Assume typical unless deductions are warranted)
Fireplaces	Number of fireplaces
FireplaceQu	Fireplace quality
GarageType	Garage location
GarageYrBlt	Year garage was built
GarageFinish	Interior finish of the garage
GarageCars	Size of garage in car capacity
GarageArea	Size of garage in square feet

GarageQual	Garage quality
GarageCond	Garage condition
WoodDeckSF	Wood deck area in square feet
OpenPorchSF	Open porch area in square feet
EnclosedPorch	Enclosed porch area in square feet
3SsnPorch	Three season porch area in square feet
ScreenPorch	Screen porch area in square feet
PoolArea	Pool area in square feet
PoolQC	Pool quality
Fence	Fence quality
MiscFeature	Miscellaneous feature not covered in other categories
MiscVal	$Value of miscellaneous feature
MoSold	Month Sold (MM)
YrSold	Year Sold (YYYY)
SaleType	Type of sale
SaleCondition	Condition of sale

4. Null values present in:

```
MasVnrType         7
BsmtQual          30
BsmtCond          30
BsmtExposure      31
BsmtFinType1      30
BsmtFinType2      31
FireplaceQu      551
GarageType        64
GarageFinish      64
GarageQual        64
GarageCond        64
LotFrontage      214
MasVnrArea         7
GarageYrBlt       64
```

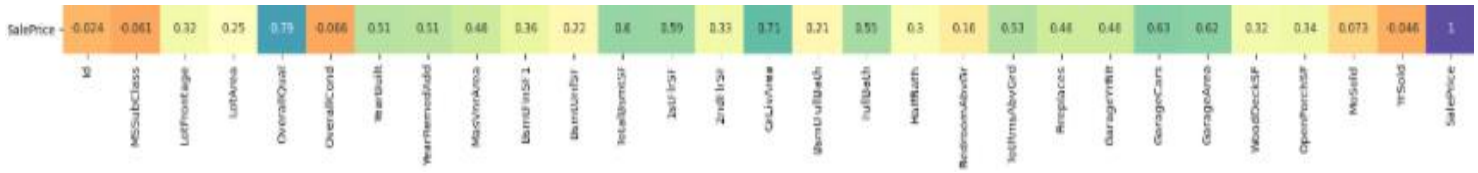
Data Pre-processing Done

1. Dropped these ['Alley','PoolQC','MiscFeature','Fence'] as they had more than 80% null values and Id which was unique for each data point.
2. Label encoding the categorical features with dtype = ‘object’

3. Imputing null values with Mode in categorical columns.
4. Imputing null values with mean in numerical columns.
5. PCA reduced the final cleaned features from 43 to 32.
6. Repeated all the steps in test dataset.
7. Outliers cannot be removed as the loss in data was massive – around 25%

Data Inputs- Logic- Output Relationships

Num_df dependency on target variable



Cat_df correlation with target variable

Highest relationship is between Sale price and Overall Quality of the house.

Hardware and Software Requirements and Tools Used

Hardware: Windows 11

Software: Jupyter notebook

Libraries: seaborn, matplotlib, statsmodels, numpy, pandas, sklearn, scipy, pickle

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

The dataset was analyzed using the standard procedure where we indulge in statistical and graphical analysis of the dataset. Statistical methods like Zscore were used but couldn't be implemented as the loss in data was huge around 25% which is not feasible as data is expensive .

Best random state is found and is used to split the data to obtain an optimum score.

Regression algorithms were used to create a model as the target variable was a continuous variable.

Testing of Identified Approaches (Algorithms)

1. Lasso Regression
2. Decision Tree Regressor
3. Random Forest Regressor
4. XGB regressor

Run and evaluate selected models

```
ls = Lasso(alpha=10,random_state=0)
ls.fit(x_train,y_train)
ls.score(x_train,y_train)
pred_ls = ls.predict(x_test)
```

```
lss = r2_score(y_test,pred_ls)

for k in range(2,10):
    lsscore=cross_val_score(ls,x,y,cv=k)
    lrcv=lsscore.mean()
    print("At cv= ",k)
    print("Cross Val score : ",lrcv*100)
    print("r2 score is : ",lss*100)
    print("\n")
```

```
dtr = DecisionTreeRegressor(criterion='poisson',
                            max_depth= 8,
                            max_leaf_nodes= 10,
                            min_samples_leaf= 100,
                            min_samples_split= 10)
dtr.fit(x_train, y_train)
pred_dtr= dtr.predict(x_test)
```

```
dtrr2 = r2_score(y_test,pred_dtr)
from sklearn.model_selection import cross_val_score
for k in range(2,10):
    dtrscore=cross_val_score(dtr,x,y,cv=k)
    dtrcv=dtrscore.mean()
    print("At cv= ",k)
    print("Cross Val score : ",dtrcv*100)
    print("r2 score is : ",dtrr2*100)
    print("\n")
```

```
rfr = RandomForestRegressor(criterion='poisson',
                            max_depth= 7,
                            max_features="sqrt",
                            n_estimators=150)
rfr.fit(x_train, y_train)
pred_rfr= rfr.predict(x_test)
```

```
rfrr2 = r2_score(y_test,pred_dtr)

for k in range(2,8):
    rfrscore=cross_val_score(rfr,x,y,cv=k)
    rfrcv=rfrscore.mean()
    print("At cv= ",k)
    print("Cross Val score : ",rfrcv*100)
    print("r2 score is : ",rfrr2*100)
    print("\n")
```

```
xgb = XGBRegressor(n_estimators= 200,
                   importance_type= 'split',
                   eta=0.1, booster='dart')
xgb.fit(x_train,y_train)
xgb.score(x_train,y_train)
pred_xgb = xgb.predict(x_test)
```

```
xgbs = r2_score(y_test,pred_xgb)

for k in [2,8]:
    xgbscore=cross_val_score(xgb,x,y,cv=k)
    xgbcv=xgbscore.mean()
    print("At cv= ",k)
    print("Cross Val score : ",xgbcv*100)
    print("r2 score is : ",xgbs*100)
    print("\n")
```

r2 Score:

Model	Score
Decision Tree Regressor	70.1180692720088
Random Forest Regressor	70.1180692720088
Lasso Regression	86.07869890407422
XGB Regressor	86.82333863575415

Cross Validation Score:

Model	Score
Decision Tree Regressor	69.11456879963502
Random Forest Regressor	17.611025661927858
Lasso Regression	78.8602883531012
XGB Regressor	86.23476477602651

Key Metrics for success in solving problem under consideration

The least difference between r2score and cv score was seen in gradient boosting regressor hence proving least fitting problem. Hence it is considered for modelling the data under consideration (test data).

The r2 score was used.

R-squared is a metric of correlation. Correlation is measured by “r” and it tells us how strongly two variables can be related.

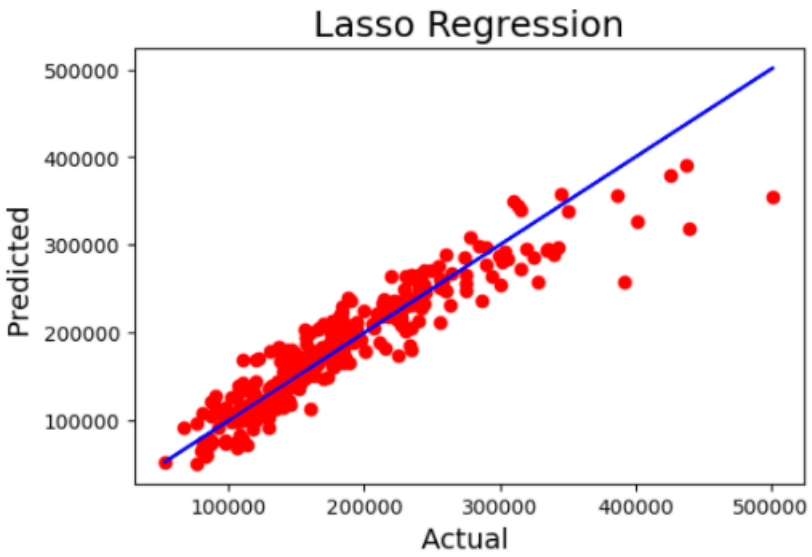
A correlation closer to +1 means a strong relationship in the positive direction, while -1 means a stronger relationship in the opposite direction.

A value closer to 0 means that there is not much of a relationship between the variables. R-squared is closely related to correlation

1. Lasso Regression

```
print('MAE:', metrics.mean_absolute_error(y_test, pred_ls))
print('MSE:', metrics.mean_squared_error(y_test, pred_ls))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, pred_ls)))
```

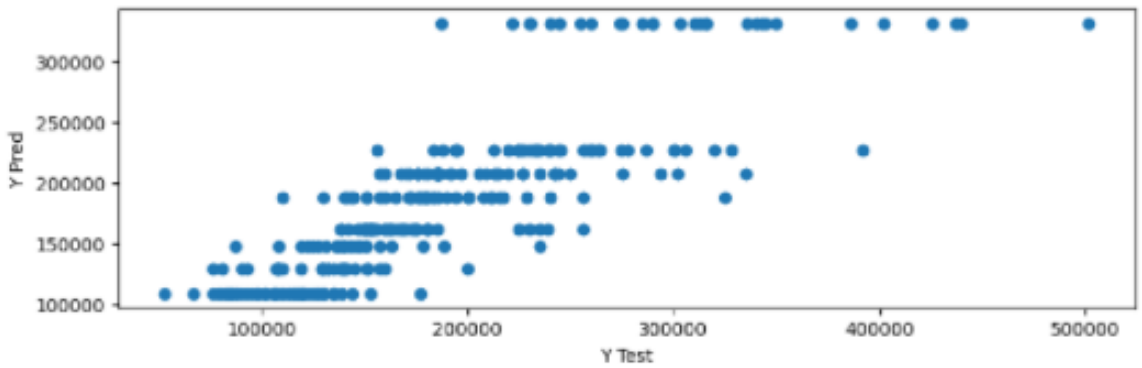
```
MAE: 20465.160043436088
MSE: 757807007.5203252
RMSE: 27528.29467148892
```



2. Decision Tree Regressor

```
print('MAE:', metrics.mean_absolute_error(y_test, pred_dtr))
print('MSE:', metrics.mean_squared_error(y_test, pred_dtr))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, pred_dtr)))
```

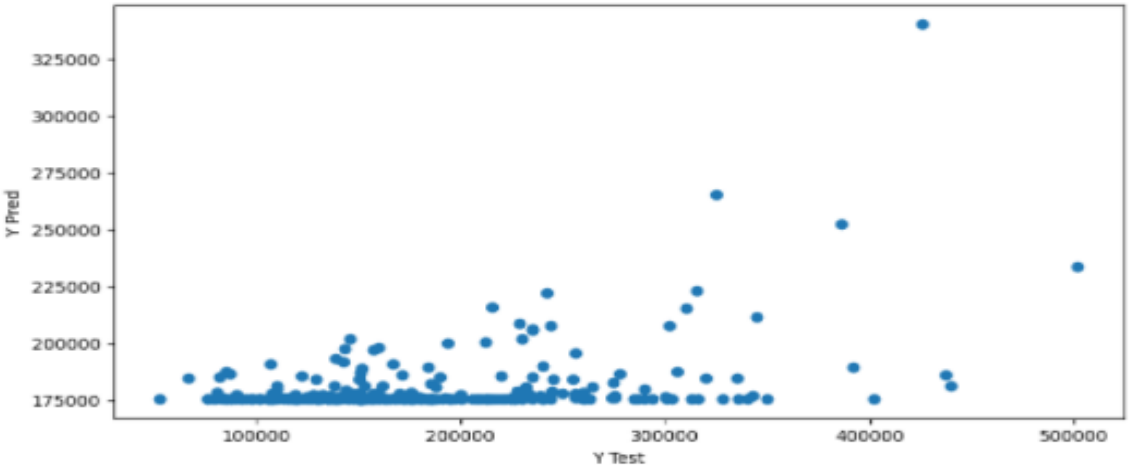
```
MAE: 27980.693527848263
MSE: 1626625007.8116531
RMSE: 40331.4394463135
```



3. Random Forest Regressor

```
print('MAE:', metrics.mean_absolute_error(y_test, pred_rfr))
print('MSE:', metrics.mean_squared_error(y_test, pred_rfr))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, pred_rfr)))
```

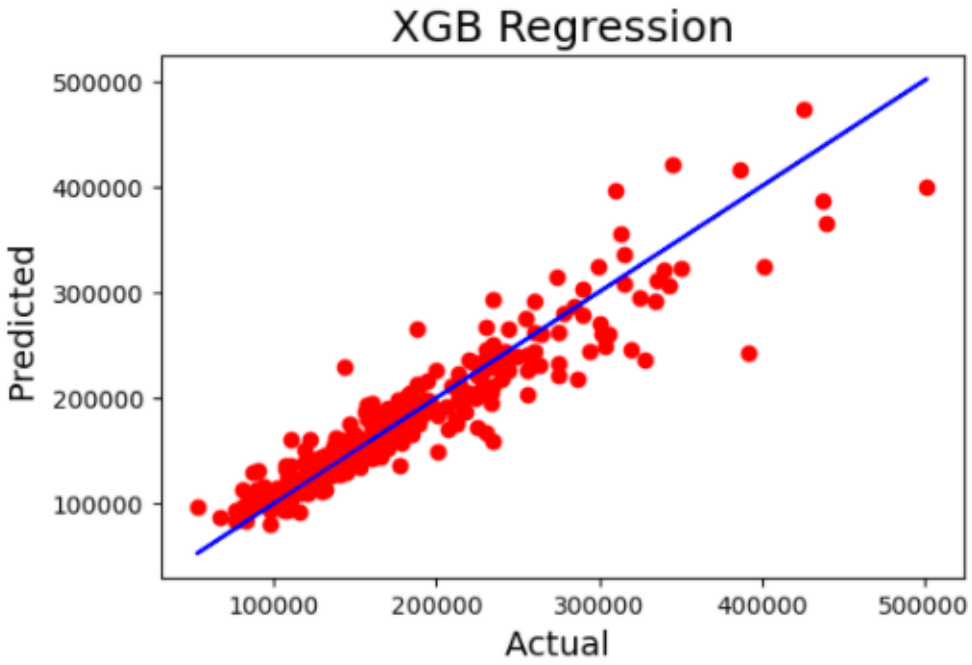
```
MAE: 53399.61528861743
MSE: 4787141941.671064
RMSE: 69189.17503245045
```



4. XGB Regressor

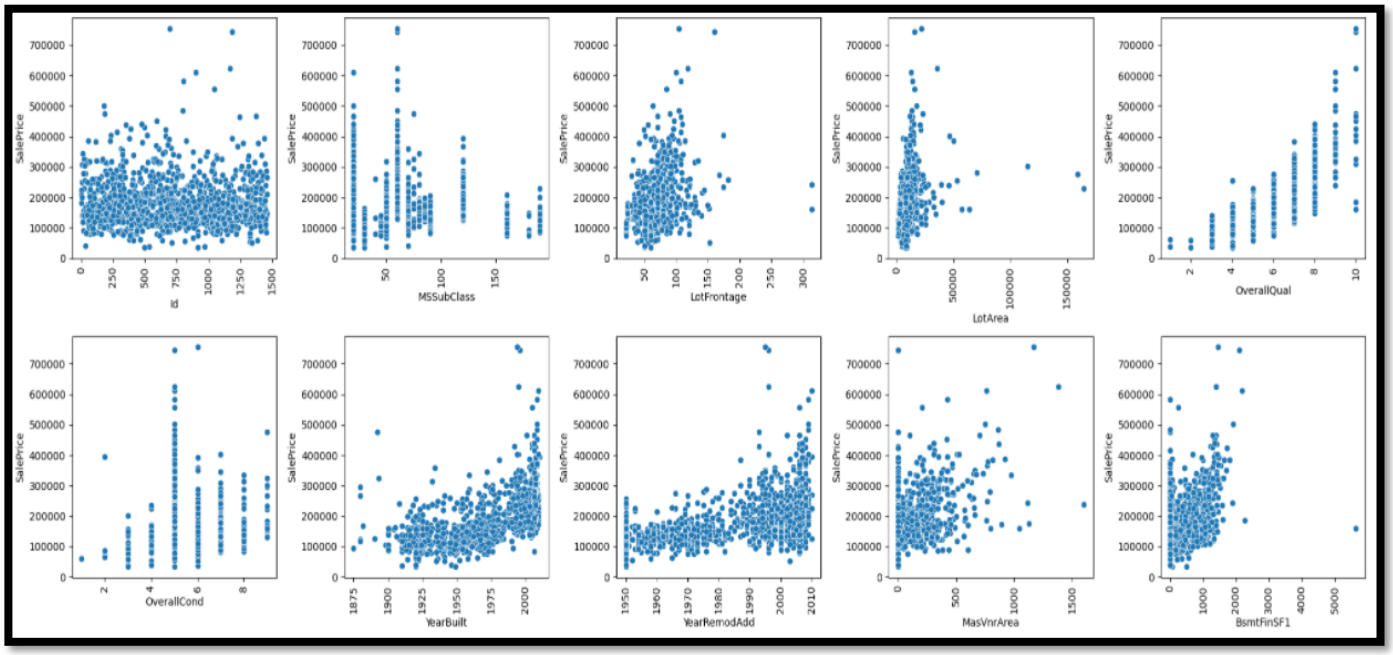
```
print('MAE:', metrics.mean_absolute_error(y_test, pred_xgb))
print('MSE:', metrics.mean_squared_error(y_test, pred_xgb))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, pred_xgb)))
```

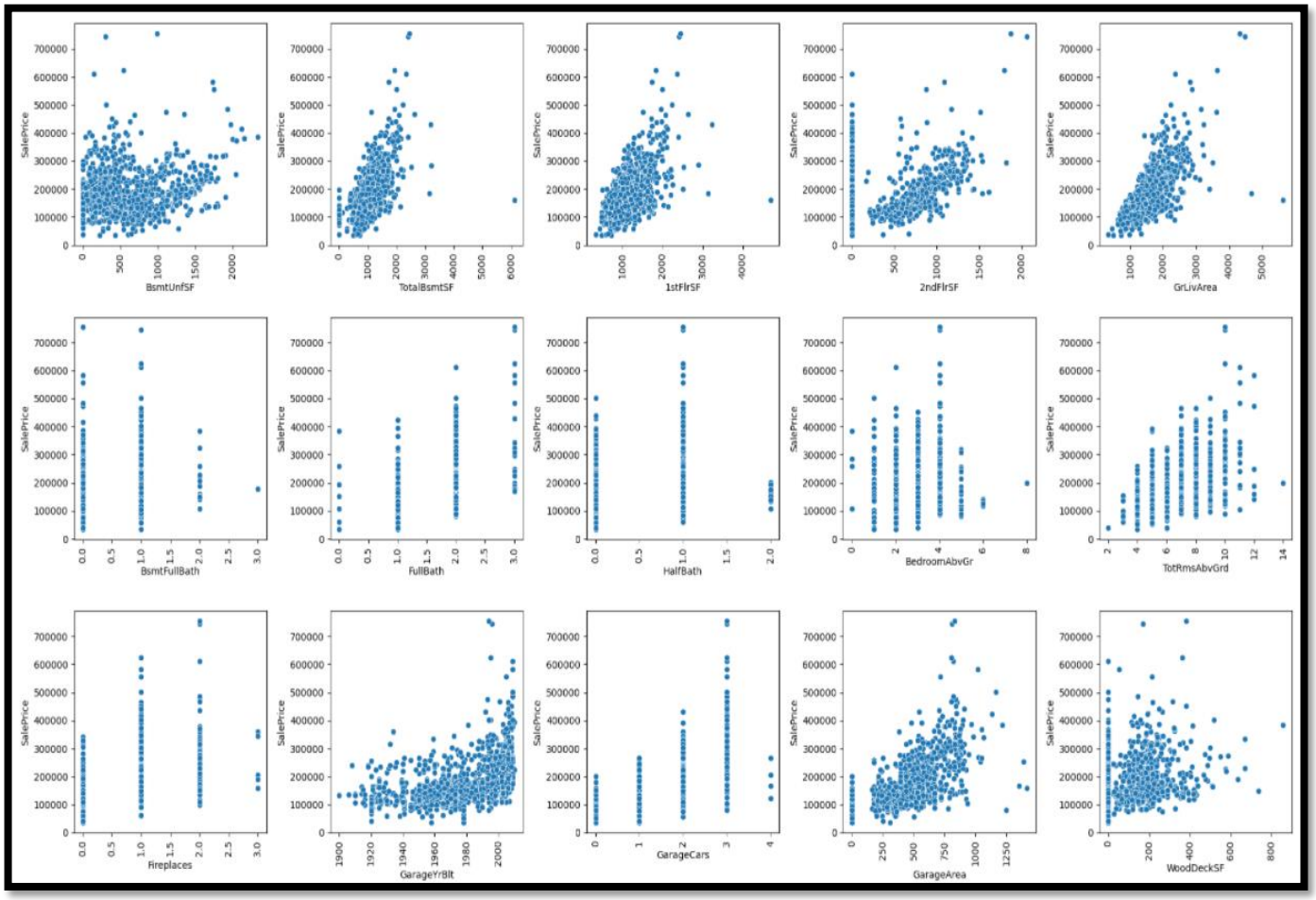
```
MAE: 18401.17053724315
MSE: 717272491.1804504
RMSE: 26781.94337945718
```



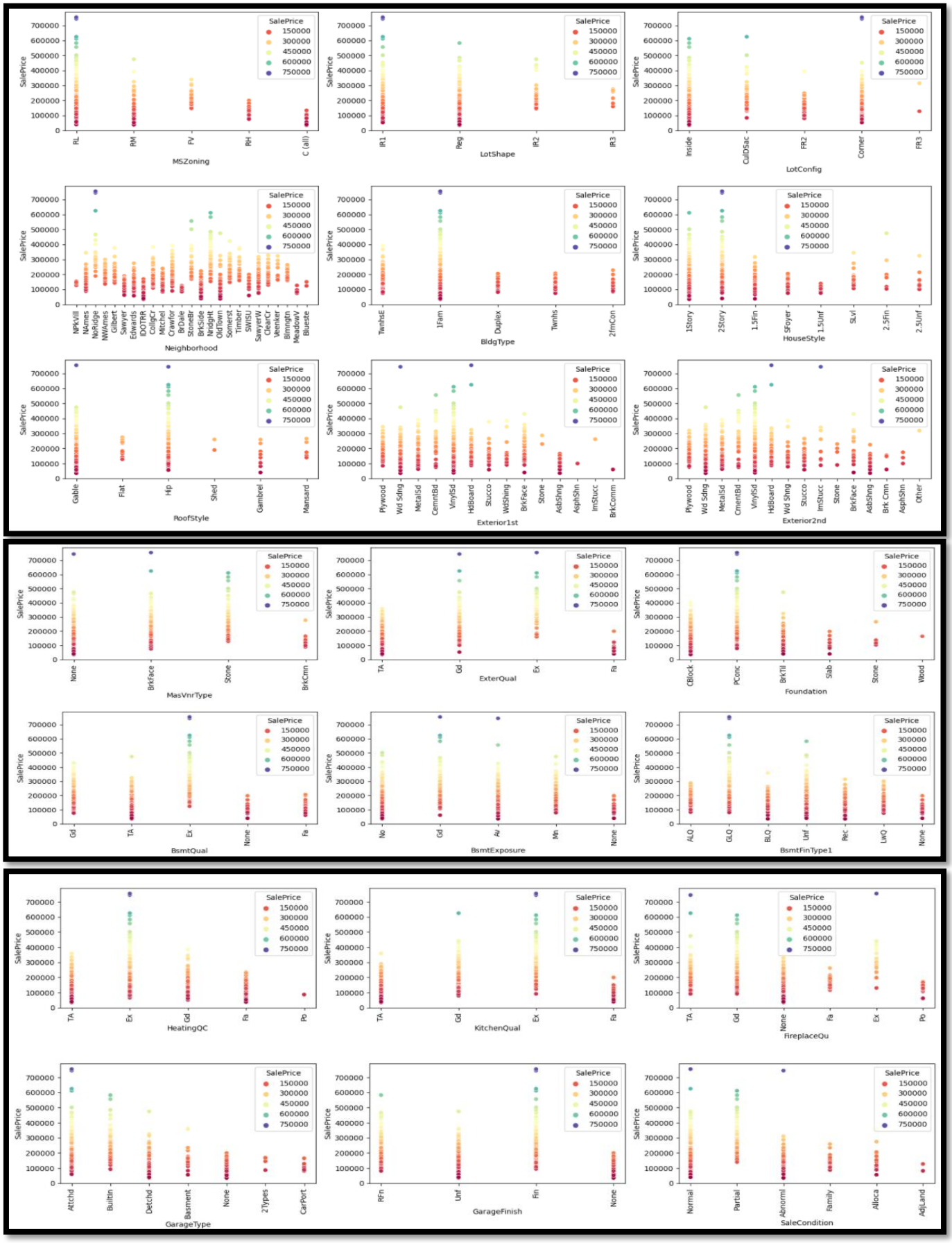
Visualizations

- Num_df plotted against Sale Price

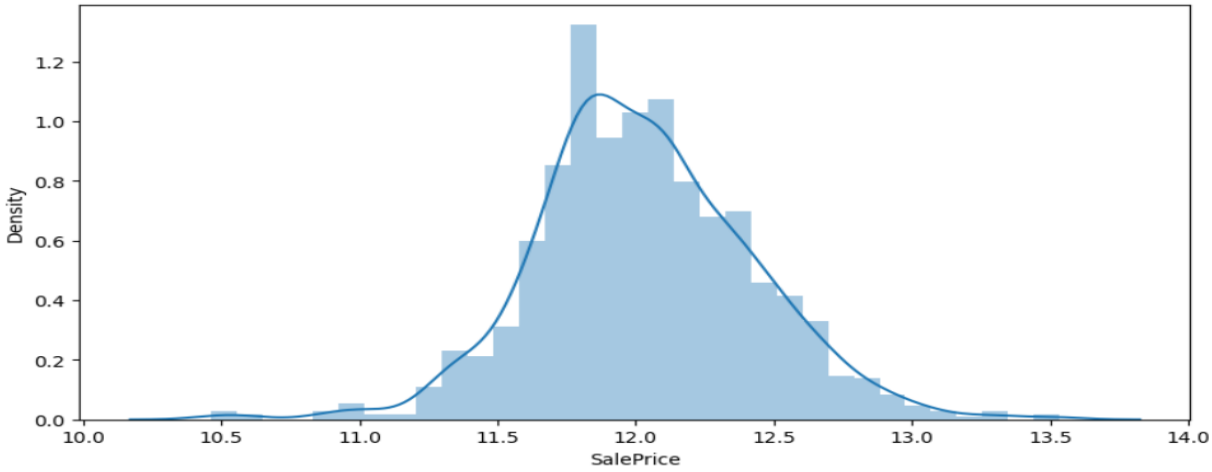




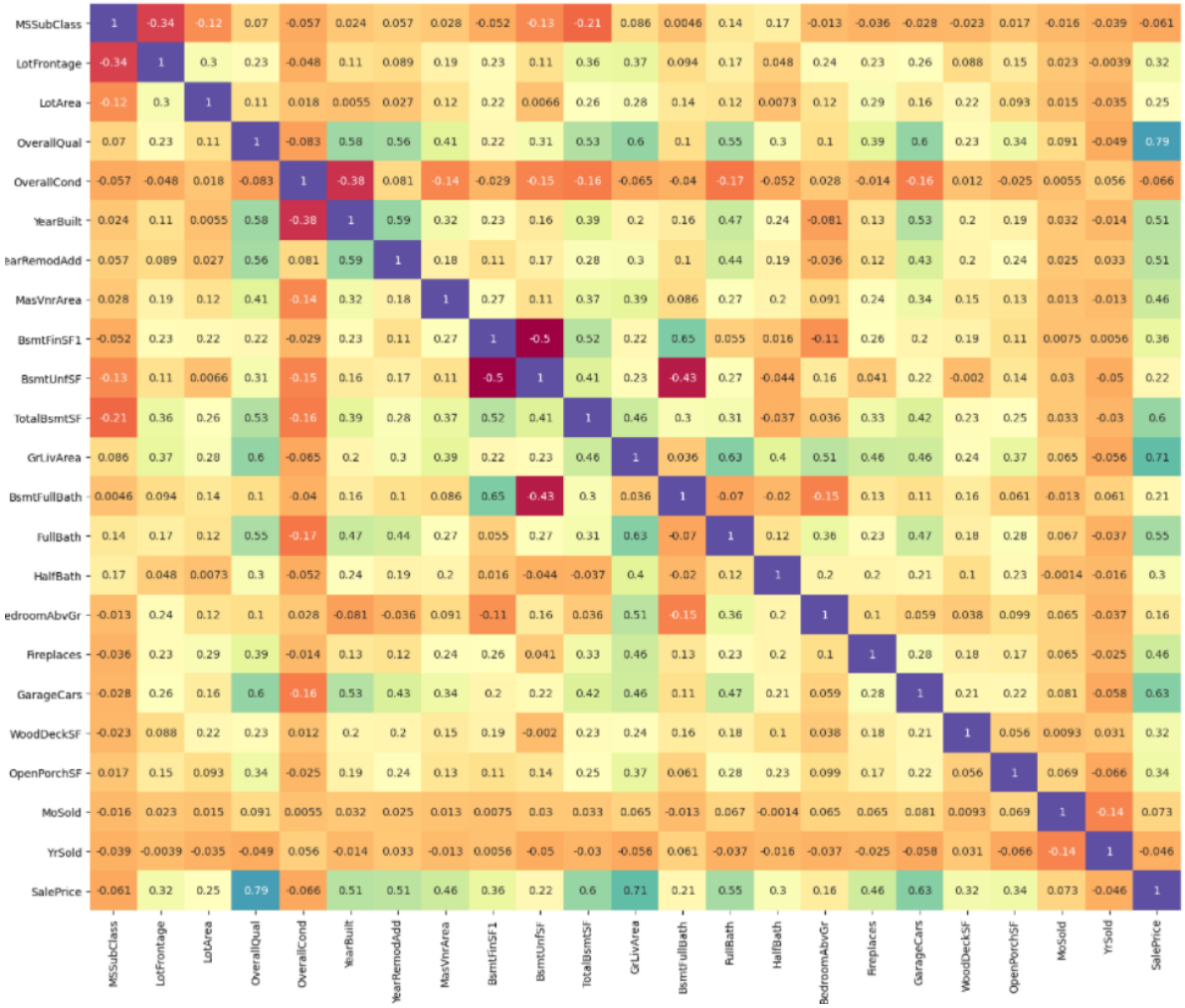
- Cat_df plotted against target column



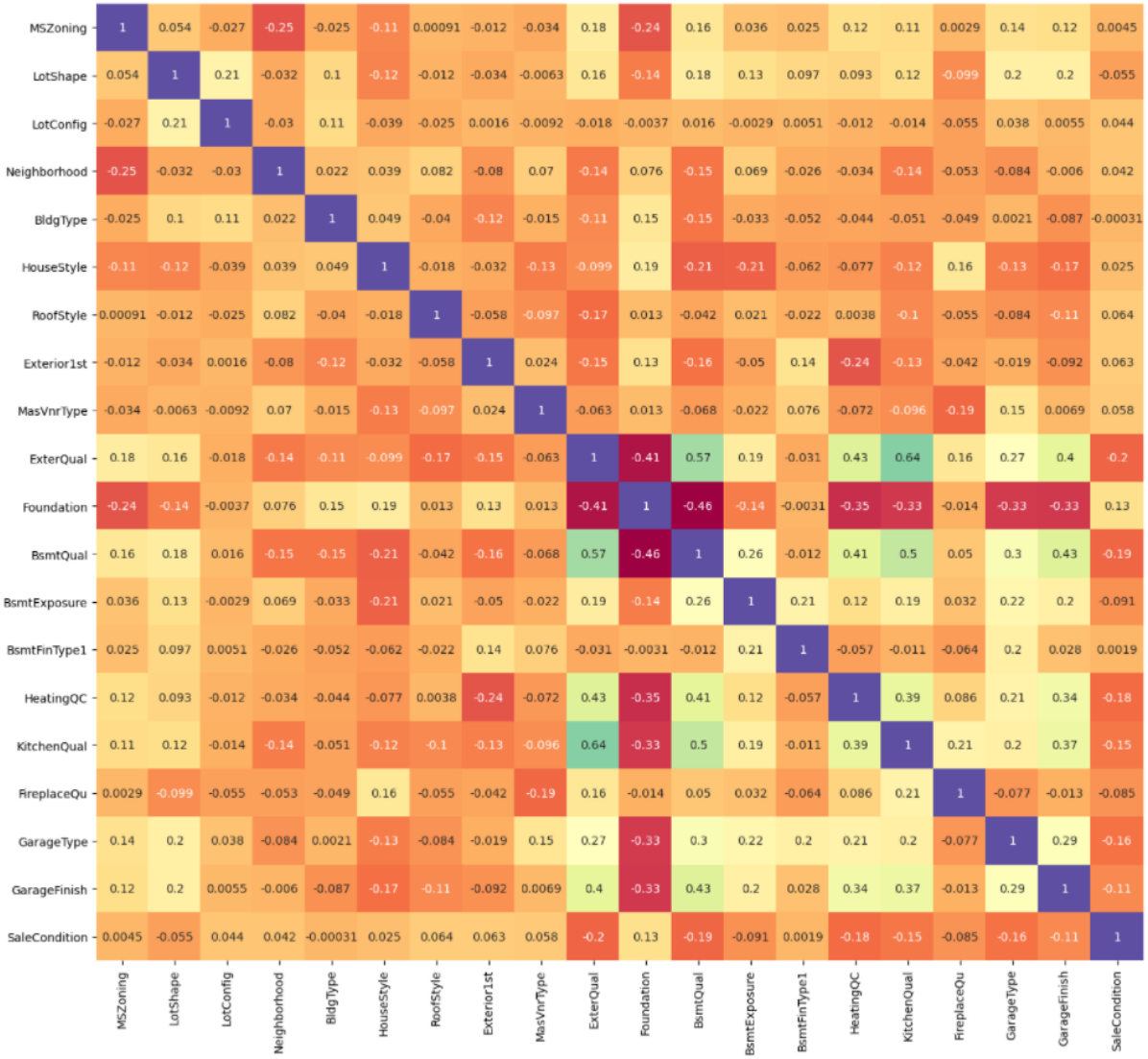
- Target variable's distribution:



- Num_df corr. heatmap (after removing multi-collinear features)



- Cat_df corr. heatmap (after removing multi-collinear features)



CONCLUSION

Key Findings and Conclusions of the Study

- Max. houses are single family detached but max sale price is for Townhouse end unit.
- Majority of houses are 1 story but sale price is the highest for 2 storey houses.
- Max houses have regular shape but slightly irregular shows the highest price.
- Max houses have overall quality as above average.

- Most house are in average overall condition but prices are increasing till average but it drops and plateaus afterwards with a slight increase.
- Majority of houses have an inside Lot but Cul de Sac have the highest prices as it provides more security and a closed sense of community.
- Houses located on Hillside with slope from side to side have the highest price due to its location but majority of the houses were a levelled ground.
- Houses with severe slopes attracts the lowest prices and vice versa.
- Residential Low density and medium density bags the highest price.
- Max houses have no exposure to walkout, but houses with good exposure to such lands the highest price.
- Max houses are having wood deck area under 200
- Most houses have paved road access.
- Sale price is the highest for new houses logically and most houses were built after 2000
- Linear relationship between year remodelled and sale price with later 2000's having the higher prices.
- Max roofs are gable and hip and also the highest sale price
- Most houses have Standard composite shingle roofs which also fetches the higher prices but Wood shakes have the highest prices.
- Vinyl Siding has the highest price
- Houses with Exterior Quality as Excellent have the highest prices even though they are a minority. Majority belongs to average house.
- Good ratings for exterior material fetch the highest price.
- Majority of houses have cinder block and poured concrete and also have the highest prices.
- Majority of houses have unfinished basements, but max sale price is for houses with good living quarter basements.
- Sale prices decreases with an increase in unfinished basement area.
- Sale price increase when total basement area increases.
- Sale condition was partial which had the max sale price

- Max houses were sold as warranty deed-conventional and max sale price is for newly constructed houses and immediately sold.
- 2007 is when the sale price was the highest. This was before the recession of 2008 and the real estate prices crashed everywhere and are now able to reach the same peak.
- Sale price is highest for houses whose garage can accommodate 3 cars but houses with garage capacity of 2 were highest in quantity.
- Max houses have unfinished garages but price is high for finished.
- Houses with 1 or 2 fireplaces have the highest price
- Max Sale Price is for houses with 2 full bathrooms above grade
- Maximum houses have 0 half baths
- Max number of houses have 3 bedrooms above ground but the sale price is highest for houses with 4 rooms above ground

Learning Outcomes of the Study in respect of Data Science

This dataset sheds light on the power of visualization and EDA. As the data is largely limited which makes it expectedly inconclusive to perform exploratory data analysis and reach any fruitful conclusions. This is first opportunity I have received to understand such a project which can directly have a real world impact through machine learning.

Limitations of this work and Scope for Future Work

The dataset is severely limited and can be expanded extensively with data mining. The number of datapoints is very low. To have an accurate model, more data needs to be mined or obtained from the company and cleaned with values of all the features available rather entering 0's and null values. Data can become highly multicollinear which can become a hurdle.