House Price Prediction & Mortgage Approval Prediction

CSCI-6660-01-Introduction to Artificial Intelligence

Tharangini

Tunga Dilip Kumar

Content

- Introduction
- Overview
- Algorithms Used
- Models
  - House Price model
  - Mortgage model
- Future Outcomes
- Conclusion
- References

Introduction

- House price forecasting and mortgage approval prediction are the two most important factors of real estate. This literature attempts to determine valuable information from various data collected property markets.
- Machine learning techniques are applied to analyze historical property transactions in United states to develop useful models for house buyers and sellers.
- Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the cities of United states.
- Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is a competitive approach.

- A mortgage, also referred to as a mortgage loan, is an agreement between the borrower and a mortgage lender to buy or refinance a home without having all the cash upfront.
- In America, 30 percentage of mortgage applications are declined and 60% of all American homes are under mortgage.
- House price data reached an all-time high of 18.4% in mid 2021,which forces majority of the people to buy the houses in mortgage.
- Majority of the people didn’t know whether they are eligible for mortgage which leads to contacting untrust-worthy brokers and wasting resources and time.
- Our model will help people to know whether they are eligible for mortgage or not.

Overview

- The ability to precisely classify observations is extremely valuable for predicting the house price and whether a particular user will be eligible for loan or not.
- We have used two models to predict the house price details and the mortgage details. They are

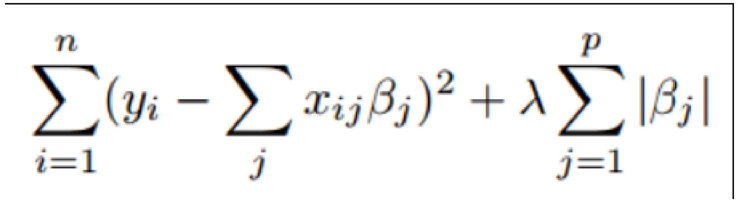
a. Lasso Regression – used for house price prediction.

b. Random Forest Classifier – used for mortgage approval prediction.

Algorithms Used

Lasso Regression:

- Lasso regression is a commonly used type of predictive analysis.
- Lasso regression is a type of linear regression that uses shrinkage it will shrink data towards the mean it is better when we had multi collinearity and to deal with selected values or to deal with certain parameters.
- Cost function for Lasso regression



$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Here λ is called turning factor or shrinkage factor
- When λ=0 then all variables are equivalent, or no shrinkage is needed

Random Forest Classifier :

- A random forest is a machine learning technique that’s used to solve regression and classification problems.
- It is the process that combines many classifiers to provide solutions to complex problems.
- A random forest algorithm consists of many decision trees. It establishes the outcome based on the predictions of the decision trees.
- It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

Use of Random Forest in Banking Sector:

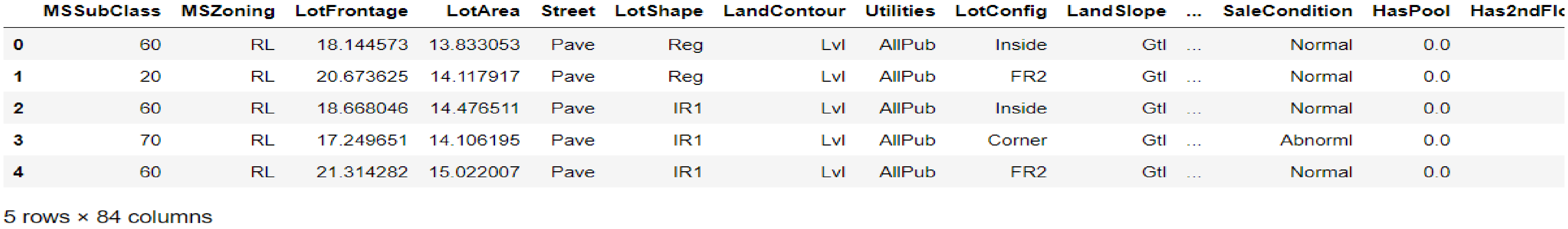
- Random forest is used in banking to predict the credit worthiness of a loan applicant.
- This helps the lending institution to make a good decision on whether to give the customer the loan or not. Banks also uses the random forest algorithm to detect fraudsters.

Models

House pricing Prediction:

- House pricing depends on multiple factors like the Home size and usable space, condition of the house, location whether it is urban, semi urban or rural and the facilities it is going to provide like kitchen, number of bedrooms and garden.
- Also factors like whether it is near the road, does the house has pool, Utilities will influence the sales price of the house.
- We use Lasso Regression model to predict the prices of houses.
- Here the data is unrefined, so we need to clean the data which is also called as data preprocessing.
- For this we need to deal with the missing values or null values then we need to deal with the string values like “Sale Condition” where there is normal or abnormal from which we can assign normal value as 1 and abnormal value as 0.

- We favor normal condition than abnormal one so we will assign weights depends to strings so that we can make every data in to a meaningful manner for data analysis.

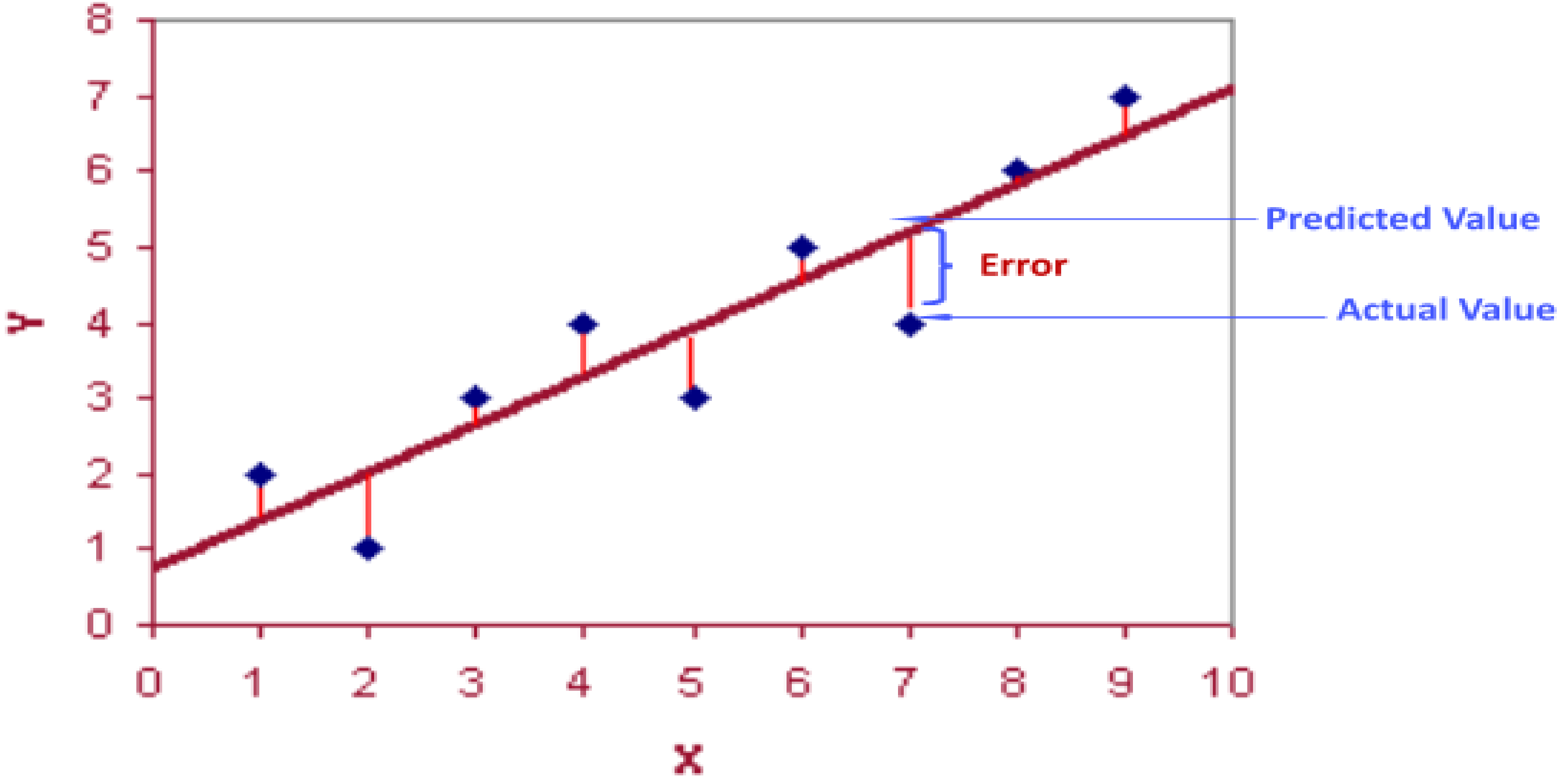


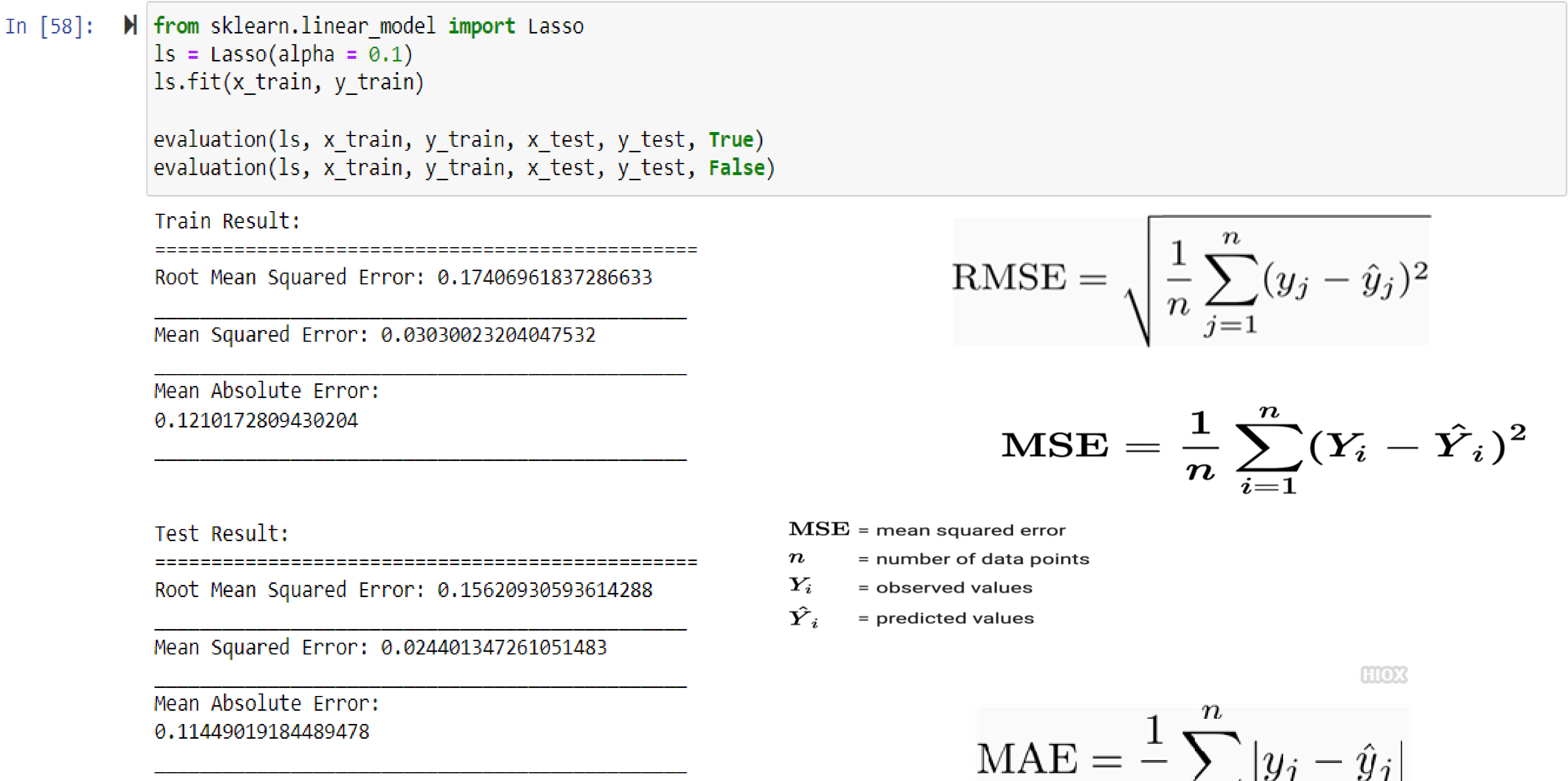
	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	Utilities	LotConfig	LandSlope	...	SaleCondition	HasPool	Has2ndFlc
0	60	RL	18.144573	13.833053	Pave	Reg	Lvl	AllPub	Inside	Gtl	...	Normal	0.0	
1	20	RL	20.673625	14.117917	Pave	Reg	Lvl	AllPub	FR2	Gtl	...	Normal	0.0	
2	60	RL	18.668046	14.476511	Pave	IR1	Lvl	AllPub	Inside	Gtl	...	Normal	0.0	
3	70	RL	17.249651	14.106195	Pave	IR1	Lvl	AllPub	Corner	Gtl	...	Abnorml	0.0	
4	60	RL	21.314282	15.022007	Pave	IR1	Lvl	AllPub	FR2	Gtl	...	Normal	0.0	

5 rows × 84 columns

- To make more sense for the data we can use matplot and seaborn to get the more effective graphical representation of data.
- As we got data in train and test we combine them initially and preprocess them and divide the data in 80 to 20 ratio
- At the final step we used lasso regression to estimate relationship between the variables and to compare the train dataset results to calculate the rmes of the model.

Graphical Representation of Linear Regression Model:





```
In [58]:  from sklearn.linear_model import Lasso
          ls = Lasso(alpha = 0.1)
          ls.fit(x_train, y_train)

          evaluation(ls, x_train, y_train, x_test, y_test, True)
          evaluation(ls, x_train, y_train, x_test, y_test, False)
```

```
Train Result:
================================================
Root Mean Squared Error: 0.17406961837286633
_______________________________________________
Mean Squared Error: 0.03030023204047532
_______________________________________________
Mean Absolute Error:
0.1210172809430204
_______________________________________________


Test Result:
================================================
Root Mean Squared Error: 0.15620930593614288
_______________________________________________
Mean Squared Error: 0.024401347261051483
_______________________________________________
Mean Absolute Error:
0.11449019184489478
_______________________________________________
```

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$\mathbf{MSE}$ = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

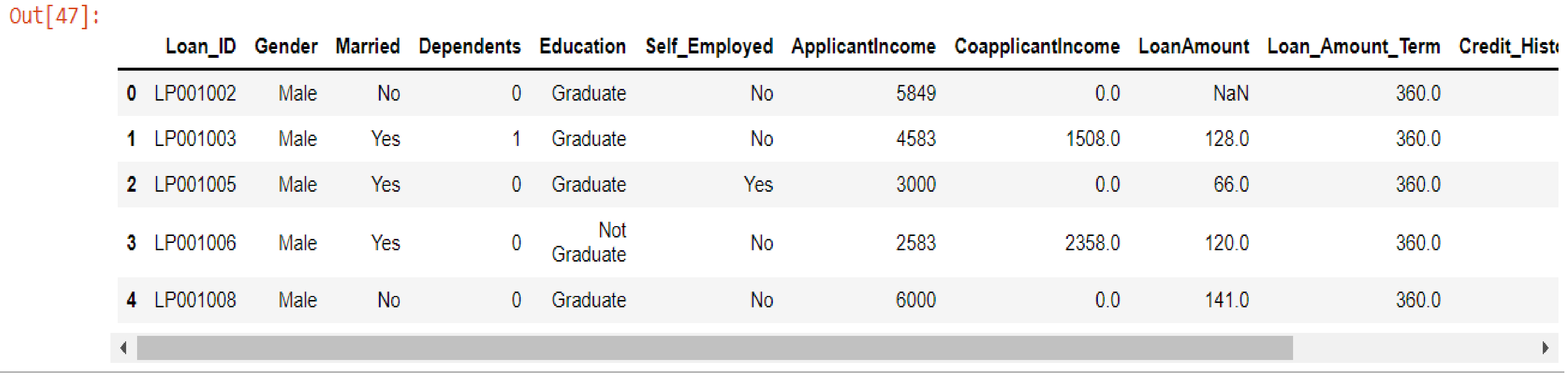
$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

©easycalculation.com

Mortgage Approval Prediction:

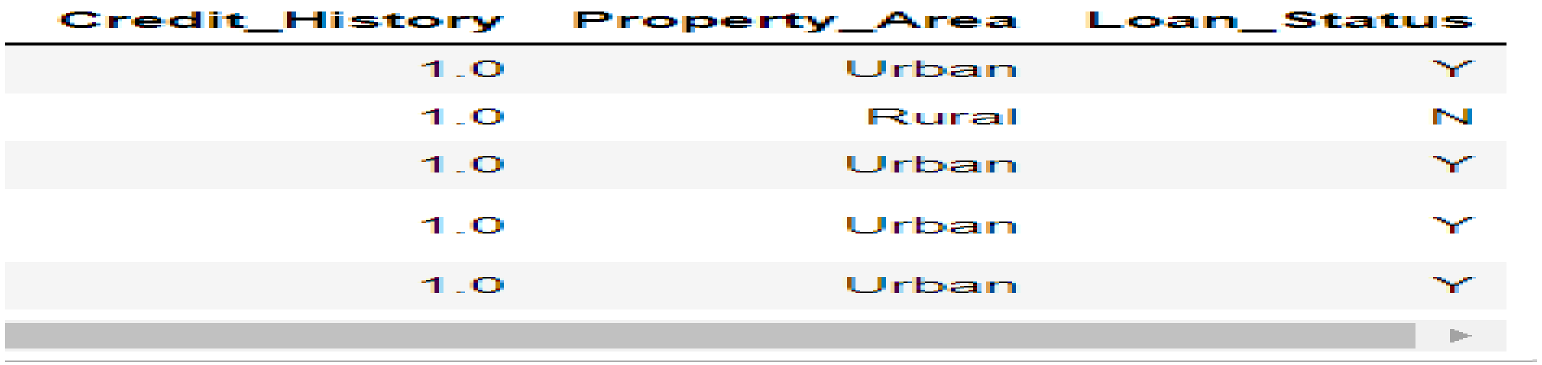
- Several factors include Gender, Education Qualification, Self Employed, Dependents and whether the person is married or not will effect the mortgage status.
- Also, the income of applicant, Loan amount, area of the property and credit history of the person are necessary to know if a certain person can get a loan or not.
- In this method, we use Random forest model to predict if mortgage application is approved or not.
- We import necessary packages and load the data set which contains data on gender employment dependents and credit history in this model.
- We can use seaborn and matplot lib to show the data in a meaningful way in the form of graphs.
- We can deal with the missing values and assign dummy values as the string values so that we can make them into single data type.

- We can also divide the dataset in to two parts which is train and test in 80 to 20 ratio and use random forest classifier to predict the accuracy of train and test model.



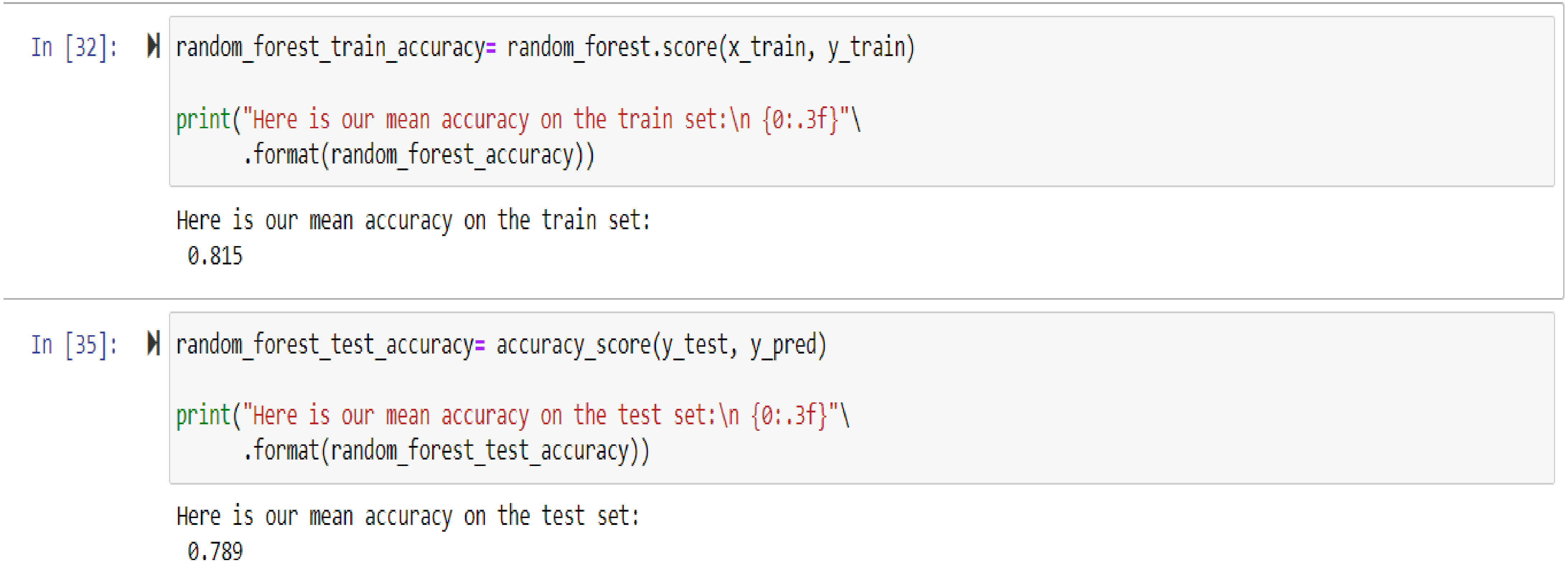
Out[47]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_Hist
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	



Credit_History	Property_Area	Loan_Status
1.0	Urban	Y
1.0	Rural	N
1.0	Urban	Y
1.0	Urban	Y
1.0	Urban	Y

- Accuracy scores for random forest classifier are



In [32]:

```
random_forest_train_accuracy= random_forest.score(x_train, y_train)

print("Here is our mean accuracy on the train set:\n {0:.3f}"\
      .format(random_forest_accuracy))
```

```
Here is our mean accuracy on the train set:
 0.815
```

In [35]:

```
random_forest_test_accuracy= accuracy_score(y_test, y_pred)

print("Here is our mean accuracy on the test set:\n {0:.3f}"\
      .format(random_forest_test_accuracy))
```

```
Here is our mean accuracy on the test set:
 0.789
```

Tools Used

Jupyter: for coding

Microsoft Excel: for data set

Packages: imported to work with data set

Pandas: for data manipulation and analysis

NumPy: for mathematical calculations

Matplotlib: for plotting graphs

Seaborn: it is used to make graph more detailed

Future Outcomes

- By using these two models end user can directly enter the data required and predict the house sale price and whether they will get a loan approved or not.

- By using them we can save lot of effort and resources as they take less effort and saves a lot of effort.

- We can further improve this model by collecting new data and changing the data feed to the model this will make it more accurate and predict more complicated situations.

Conclusion

- We can conclude that using lasso regression, we can predict the dependent variable ‘Sales price’ which is a continuous variable.

- Whereas random forest classifier will make decision trees even though it takes more space, it will provide better results.

- By using these models, we can reduce the time complexity and we can further improve this in future by adding new data to the model and make more complicated predictions.

References

- R. A. Fisher, Statistical methods for research workers, In Breakthroughs in statistics (1992), 66–70.
- Franz Fuerst and George Matysiak, Analysing the performance of nonlisted real estate funds: a panel data analysis, Applied Economics 45 (2013), no. 14, 1777–1788.
- Richard J Herring and Susan M Wachter, Real estate booms and banking busts: An international perspective, The Wharton School Research Paper (1999), no. 99-27
- Data source,
- https://www.kaggle.com/datasets/house-price-prediction?select=train.csv
- https://www.kaggle.com/datasets/house-loan-data-analysis?select=loan_data.csv