**Regression Models**

**Data Pre-processing:**

To analyze the importance of data cleaning and pre-processing we ran the sample Linear Regression model on raw data and following are the results:

Linear regression Mean Absolute Error is 566.04600684

Linear regression Root Mean Squared Error 5.67239829298

Linear regression R2 Score of the model is -236.823206277

As we can see the accuracy and error of the algorithm is very poor, we performed data pre-processing.

Handling missing/NaN values and process data:

Refer to the script Regression\_Algorithms.py, method handle\_missing\_values(df) to see how each feature has been handled. A few sample features are:

- FIRST\_PAYMENT\_DATE, MATURITY\_DATE is converted into date type

- MSA missing value is replaced with Neither MSA nor MD/Unknown

- ORIGINAL\_DEBT\_TO\_INCOME\_RATIO spaces are replaced with 66 (value greater than 65)

- ORIGINAL\_INTEREST\_RATE missing values replaced with mean value

**Feature selection:**

Since feature selection is a crucial step in building an accurate model. We used following approaches for feature selection:

* Forward
* Backward
* Stepwise

Best feature selection results were given by “Stepwise” since its AIC was maximum, thus those parameters were chosen for training the model.

**Parameter List after feature selection**:

MATURITY\_DATE

ORIGINAL\_UPB

OCCUPANCY\_STATUS

CREDIT\_SCORE

PROPERTY\_STATE

ORGINAL\_COMBINED\_LOAN\_TO\_VALUE

CHANNEL

SELLER\_NAME

PROPERTY\_TYPE

LOAN\_PURPOSE

**Linear regression:**

Used SGDRegressor with loss='epsilon\_insensitive', penalty='elasticnet'

Results:

Linear regression Mean Absolute Error is 502972351.779

Linear regression Root Mean Squared Error 5800930.01576

Linear regression R2 Score of the model is -2.48723056852e+14

**Neural Network:**

Results:

**Most optimal:**

Hidden node: [20, 10, 20]

Neural Network Mean Absolute Error is 0.120926458608

Neural Network Root Mean Squared Error 0.00151883836676

Neural Network R2 Score of the model is 0.999982949233

Hidden node: [10,20,10]

Neural Network Mean Absolute Error is 0.293112647889

Neural Network Root Mean Squared Error 0.003856006561

Neural Network R2 Score of the model is 0.999890100347

Hidden node: [30,20,30]

Neural Network Mean Absolute Error is 0.35033063229

Neural Network Root Mean Squared Error 0.00389484039889

Neural Network R2 Score of the model is 0.999887875601

**KNN:**

Results:

Mean Absolute Error is 0.0 % for K-Value: 1

Root Mean Squared Error 0.0

Accuracy of the model is 1.0

Mean Absolute Error is 0.150701687899 % for K-Value: 2

Root Mean Squared Error 0.201005038679

Accuracy of the model is 0.701369141369

Mean Absolute Error is 0.177131166386 % for K-Value: 3

Root Mean Squared Error 0.233317789474

Accuracy of the model is 0.597638423748

Mean Absolute Error is 0.188367303597 % for K-Value: 4

Root Mean Squared Error 0.248255622195

Accuracy of the model is 0.544467906251

We incremented the k value until 25 but the accuracy is getting reduced. Thus the optimal value is k=2 and results are:

Mean Absolute Error is 0.150701687899 % for K-Value: 2

Root Mean Squared Error 0.201005038679

Accuracy of the model is 0.701369141369

**Random Forest:**

**Results:**

Mean Absolute Error is 1.07616117269e-05

Root Mean Squared Error 0.000659676006272

Accuracy of the model is 0.99999678351

**Conclusion**: The performance of Random Forest algorithm is the best. However the accuracy seems to be very high.