Report

Avail Finance - Internship Assignment

By: Harsh Choudhary
IIT Kharagpur

Imputing the values

In columns having mean greater than the median, the NaN values were replaced by the median.
df['MORTDUE'],df['VALUE'],df['YOJ'],df['DEROG'],df['DELINQ'],df['CLAGE'],df['NINQ'] and df['CLNO']were imputed this way.

In columns having median greater than the mean, the NaN values were replaced by the mean.
df['DEBTINC'] was imputed this way.

In the columns having categorical data, the NaN values were replaced by the mode.
df['REASON'] and df['JOB'] were imputed this way.

One Hot Encoding

The categorical columns were one hot encoded used pandas.get_dummies.

Transforming the variables/Feature Engineering

Five new features were created :

1.)cf['new_1']=cf['DELINQ']/(cf['CLAGE']+1)

2.)cf['new_2']=(cf['DEROG']+cf['DELINQ'])/2

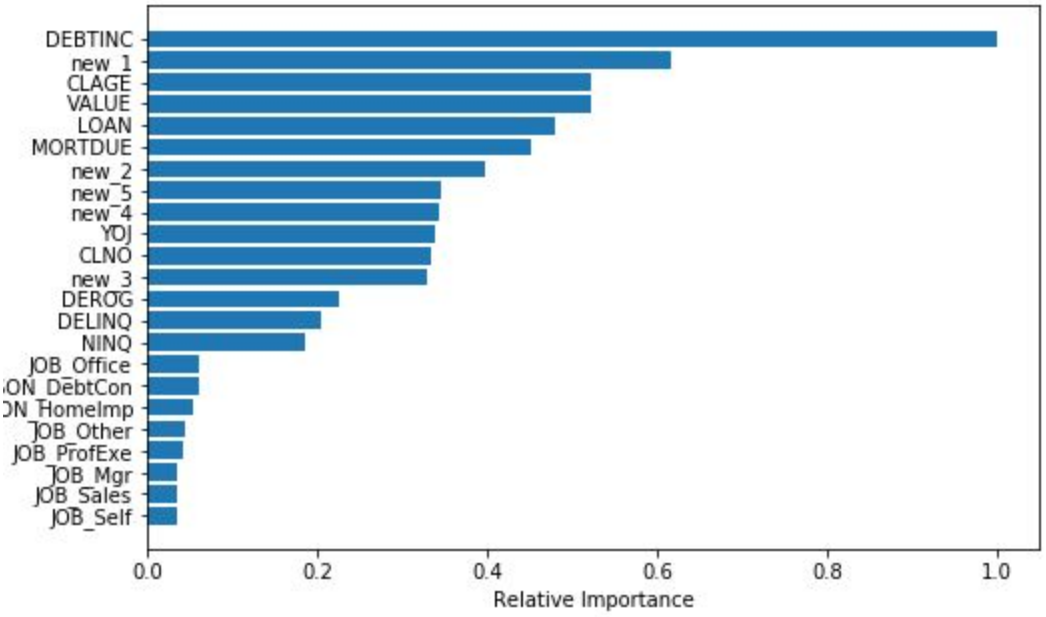
3.)cf['new_3']=cf['NINQ']/(cf['CLAGE']+1)

4.)cf['new_4']=np.sqrt(cf['YOJ']*cf['CLNO'])

5.)cf['new_5']=np.log(cf['CLNO']+1)

Feature Importance

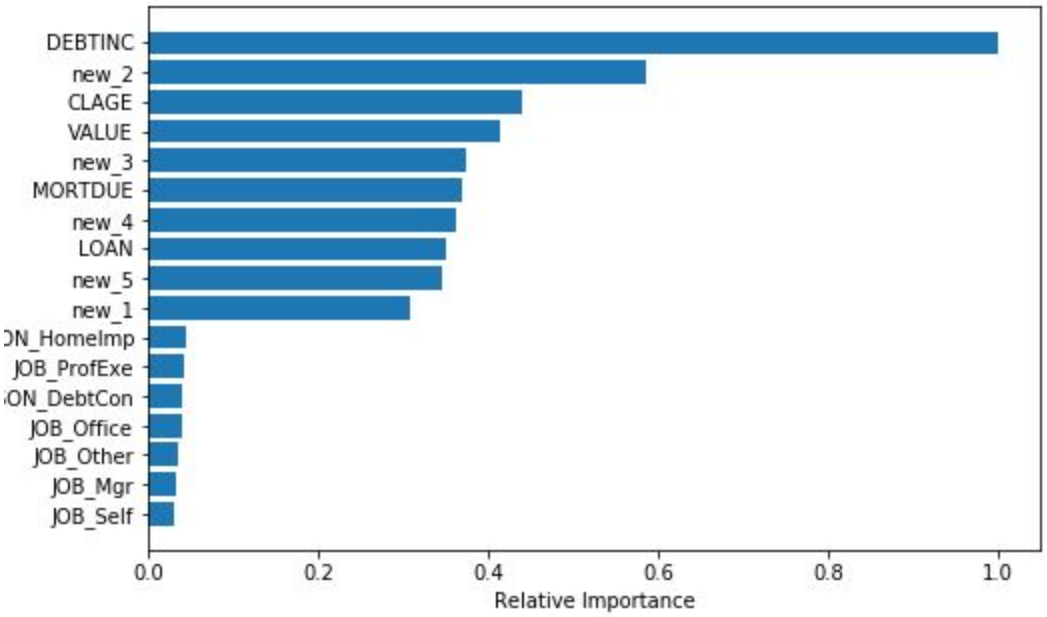
After feature engineering, the relative importance of all features was noticed was found using the decision tree.



It was found that the new features are performing better than the original features.So the corresponding original features were removed so that multicollinearity is avoided.

Variable Selection

Now the feature importance was as follows:



'DEBINC' is the most important feature.

Data Pre-Processing

Data were scaled using the MinMaxScaler and then machine learning algorithms were applied.

Modeling

The following machine learning algorithms were applied:

1.)Logistic regression
Accuracy was 0.83.

2.) Decision Tree Classifier
Accuracy was 0.88.

3.)Random Forest Classifier
Accuracy was 0.90.

4.) XGBoostClassifier
Accuracy was 0.902

So, XGBoost classifier was the top performer and was chosen for hyperparameter tuning.

Hyperparameter Tuning

Best accuracy of 0.92 was obtained with the following parameters:

learning_rate =0.05,
n_estimators=1000,
max_depth=5,
min_child_weight=2,
gamma=0.1,
subsample=0.67,
colsample_bytree=0.7,
reg_alpha=0.1,
reg_lambda=0.088,
objective= 'binary:logistic',
scale_pos_weight=1,
random_state=7,
seed=27

The Probability of loan default

By taking different probability thresholds the accuracy of the model on the test set was determined.

Maximum accuracy of 0.93 was obtained on taking the threshold as **0.218.**

So if the probability of default is more than **0.218** the application should be rejected.

Statistical Tests

Confusion matrix was used to find recall, precision and the F1 score to better understand the outcome of the model.

	precision	recall	f1-score	support
0	0.93	0.98	0.95	1575
1	0.88	0.69	0.78	392
avg / total	0.92	0.92	0.92	1967