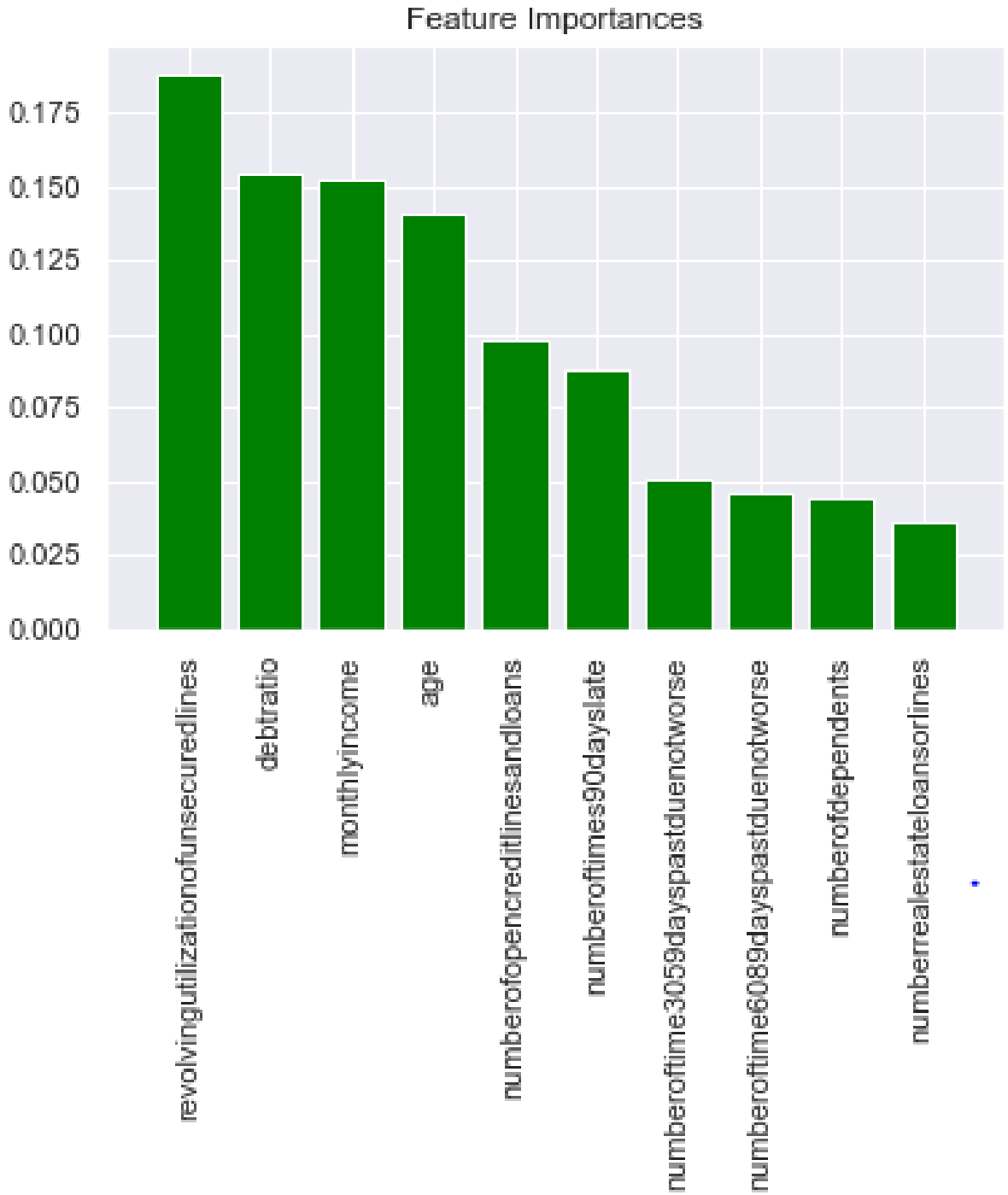
CREDIT RISK MODELING

Problem Statement

- Credit Risk refers to the chance that a customer/borrower will not to be able to make their payments on time.
- This can be modeled using various Data Models which train Machine Learning Algorithms using the data acquired from the lender.
- The main objective of these models is to predict the probability that a loan will be repaid
- Problem Statement: "If given data about a borrower, how likely is it that the loan taken will be repaid?"

Data Preparation:

- Column names were cleaned
- Missing data was replaced with median values of each column
- Outliers were dealt with by looking at the value counts of each value in that column using the 'Counter' function



- The importance of a feature in a model is found using the RandomForestClassifier.
- This gives us how importance each feature is in predicting the credit risk

Modeling:

- The following Machine Learning Algorithms will be used to model the data:
  - Logistic Regression
  - K Nearest Neighbor Classifier
  - AdaBoost Classifier
  - Gradient Boosting Classifier
  - Random Forest Classifier

We initailly split the dataset into training and testing data.

For each model we create a model object, fit the training data and calculate the ROC accuracy score using the testing data

Cross Validation:

5 –fold cross validation was done on each model and here are the mean of the score values

KNeighborsClassifier’	0.593990023059820 7
'LogisticRegression’	0.848888452051404 8
'AdaBoostClassifier’	0.858608168711411
'GradientBoostingClassifier’	0.863698351489564
'RandomForestClassifier’	0.778844504439265 3

AdaBoost Classifier and **Gradient Boosting Classifier** seem to have the highest accuracy score among the other classifiers.

Best Hyper parameters

- To obtain the best hyper parameters for the above selected classifiers we use RandomizedSearchCV.
- For Ada boost the best params are {'n_estimators': 100} and the score is 0.86036
- For Gradient Boosting Classifier the best params are:
  {'loss': 'exponential', 'max_depth': 3, 'n_estimators': 205} and score is 0.864906

Final Model

- To aviod skewed results, the features can be transformed using a logarithmic function.
- Finally VotingClassifier was used to get the best model