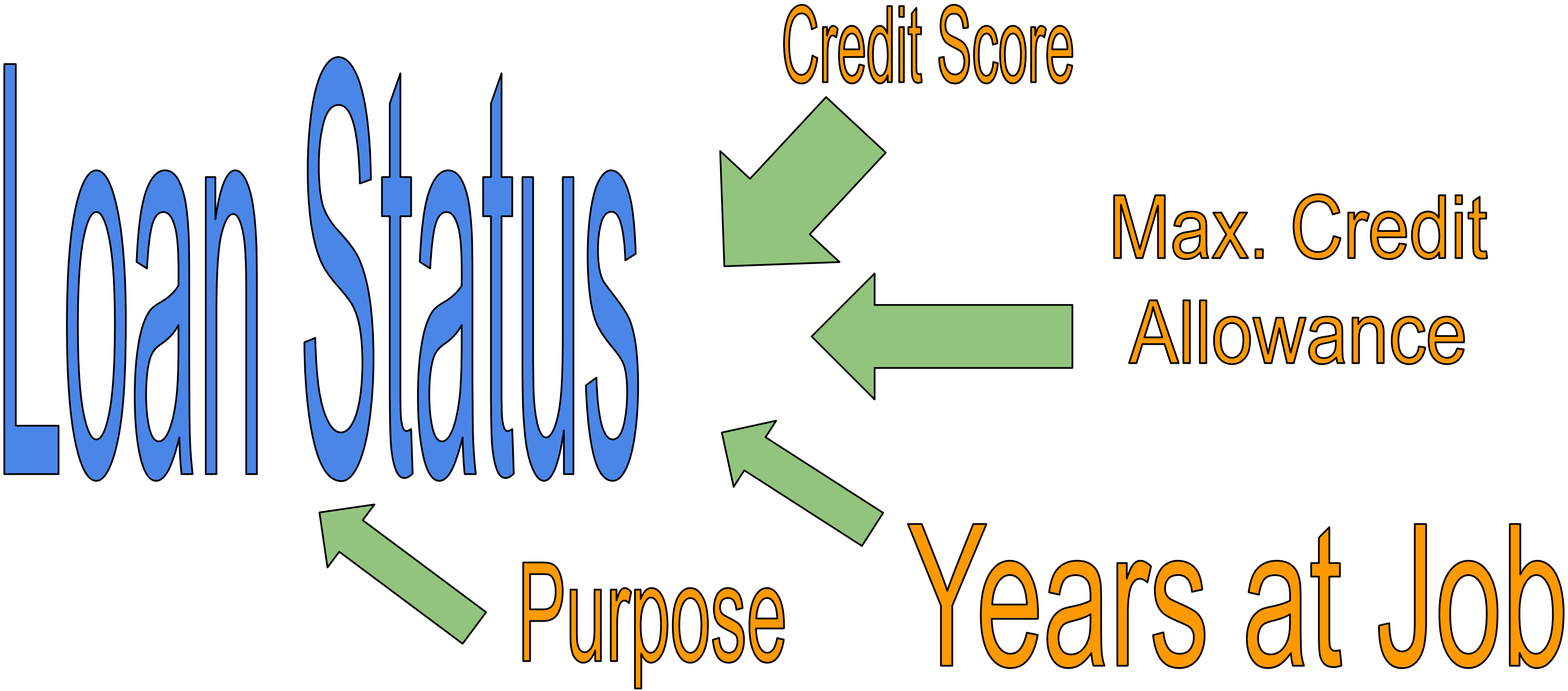
Machine Learning Applications in Finance

A presentation by Edward Beck

What are you trying to do....Predict the future, obv



Loan Status

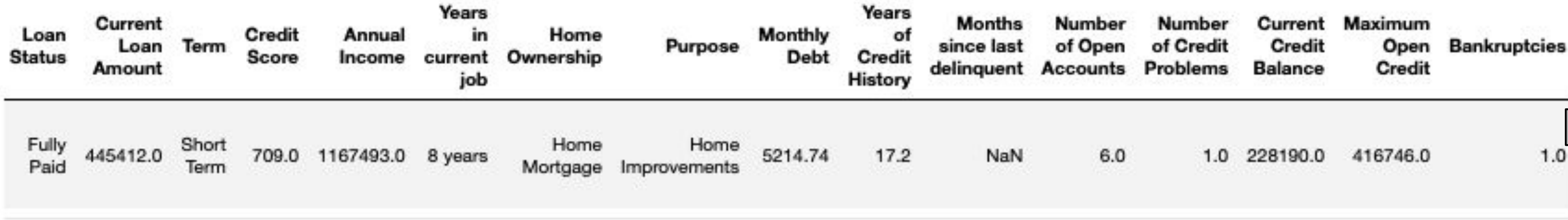
Credit Score

Max. Credit Allowance

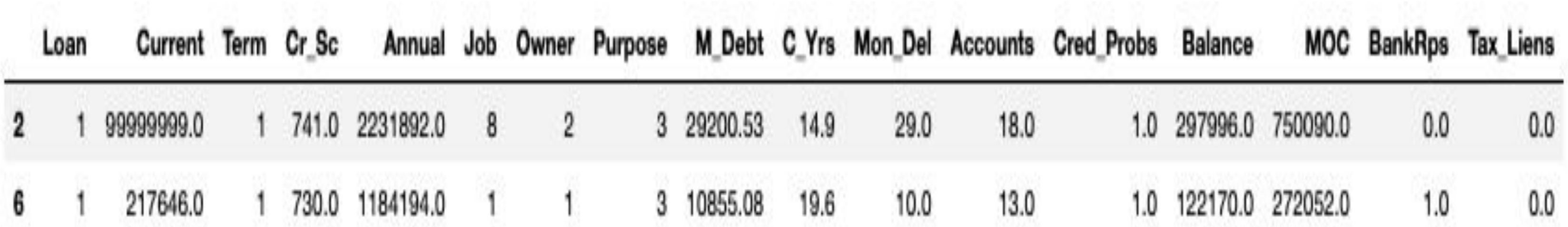
Years at Job

Purpose

The DataSet

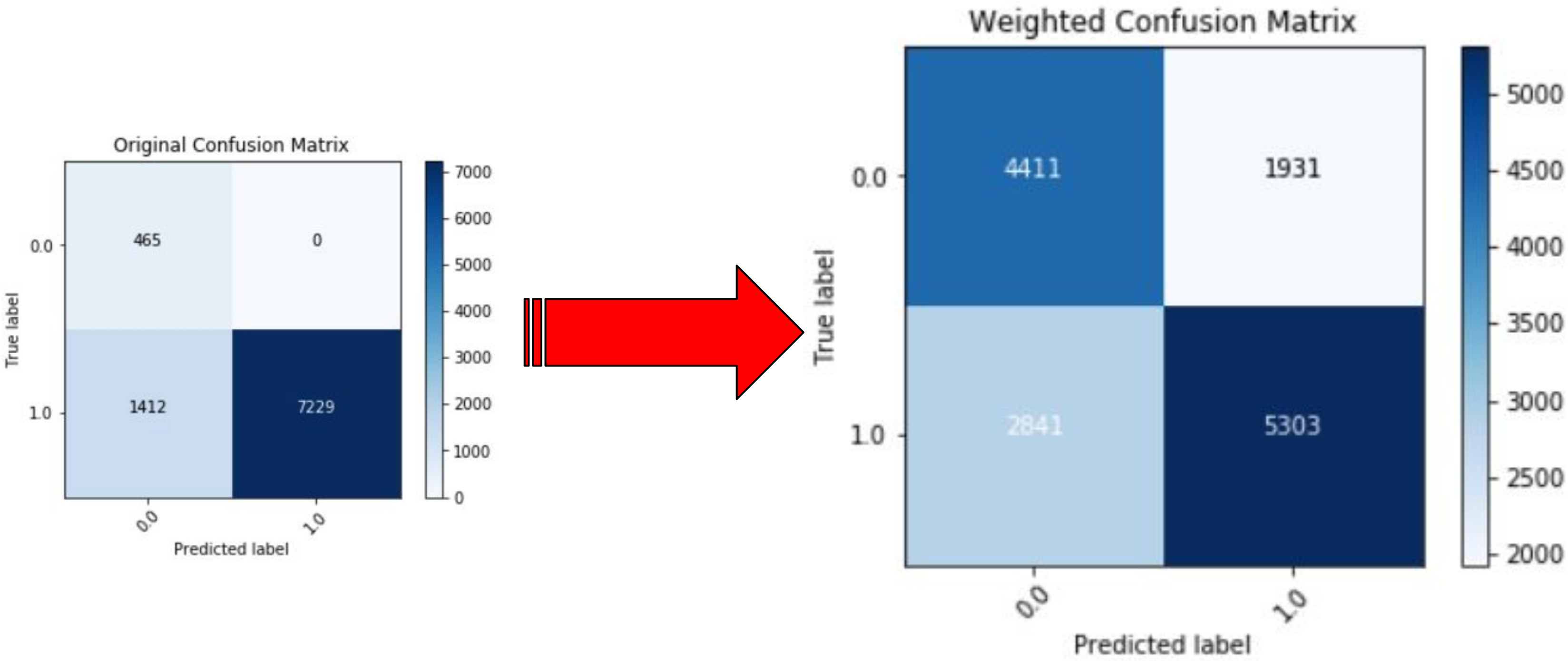


Loan Status	Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	Home Ownership	Purpose	Monthly Debt	Years of Credit History	Months since last delinquent	Number of Open Accounts	Number of Credit Problems	Current Credit Balance	Maximum Open Credit	Bankruptcies
Fully Paid	445412.0	Short Term	709.0	1167493.0	8 years	Home Mortgage	Home Improvements	5214.74	17.2	NaN	6.0	1.0	228190.0	416746.0	1.0

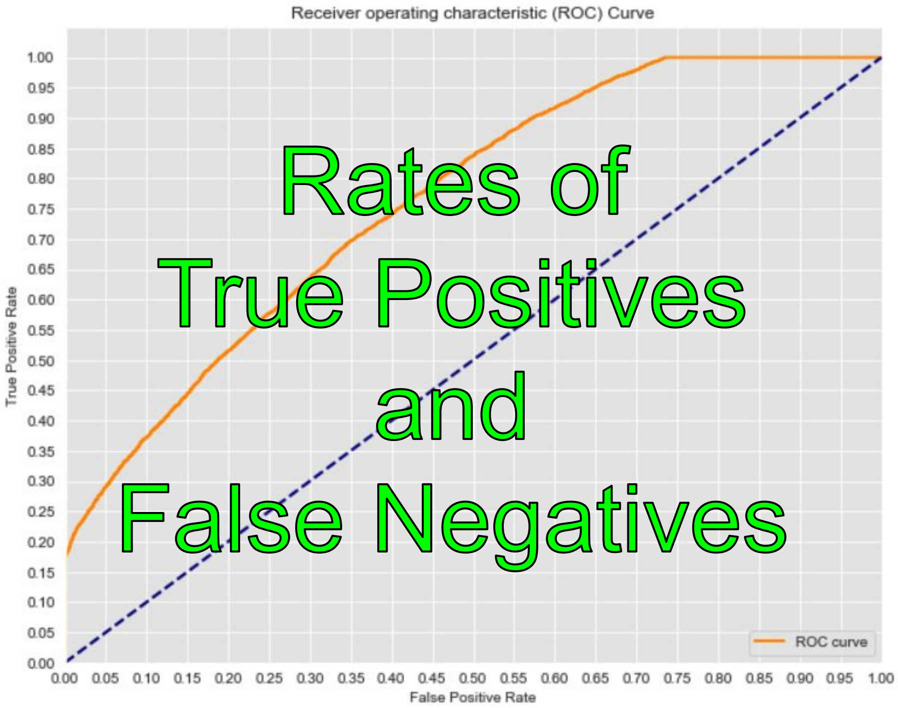


	Loan	Current	Term	Cr_Sc	Annual	Job	Owner	Purpose	M_Debt	C_Yrs	Mon_Del	Accounts	Cred_Probs	Balance	MOC	BankRps	Tax_Liens
2	1	99999999.0	1	741.0	2231892.0	8	2	3	29200.53	14.9	29.0	18.0	1.0	297996.0	750090.0	0.0	0.0
6	1	217646.0	1	730.0	1184194.0	1	1	3	10855.08	19.6	10.0	13.0	1.0	122170.0	272052.0	1.0	0.0

And some more info about the data

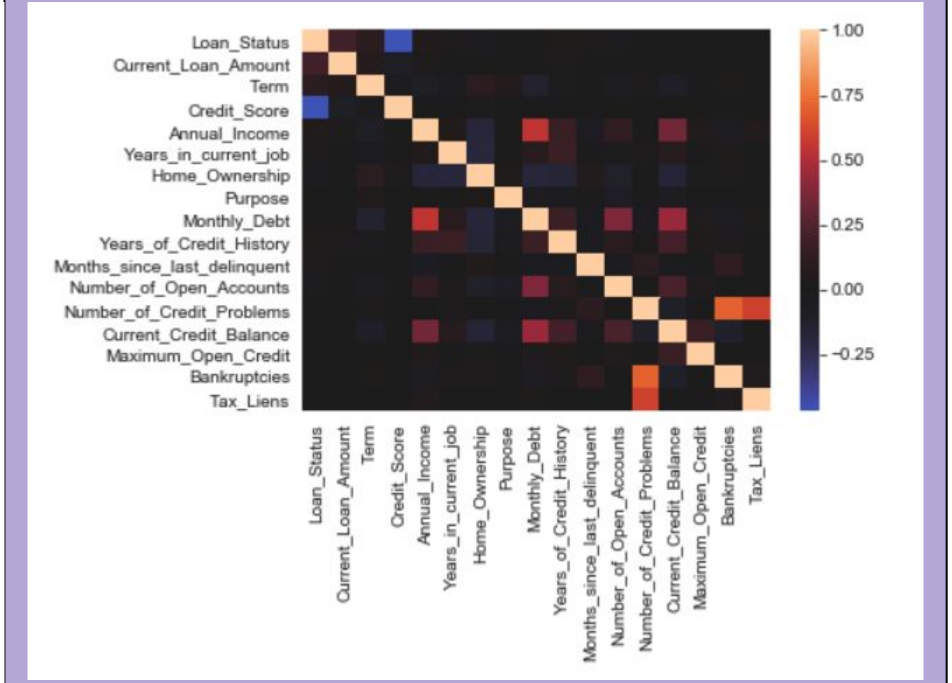


How we're going to do it.

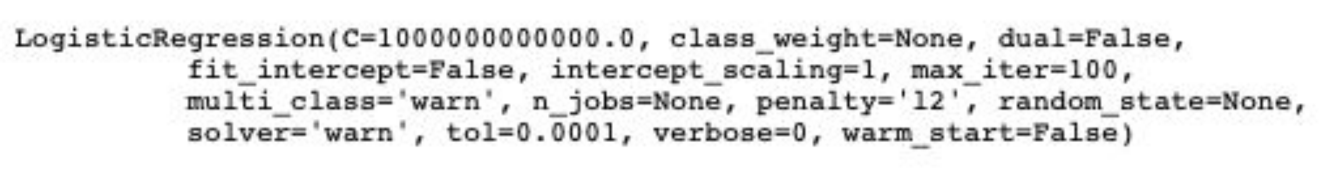


Rates of
True Positives
and
False Negatives

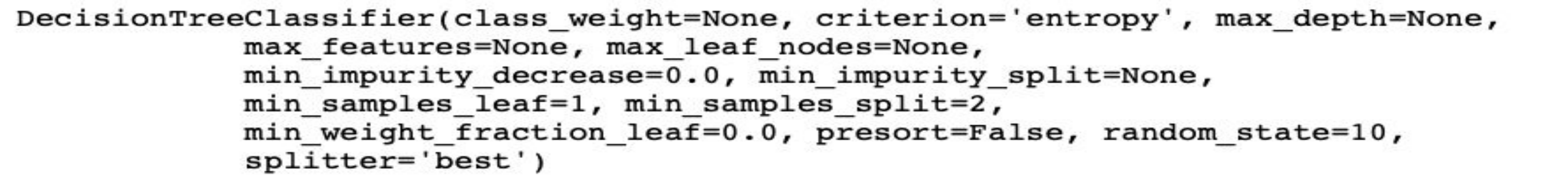
Multicollinearity Check



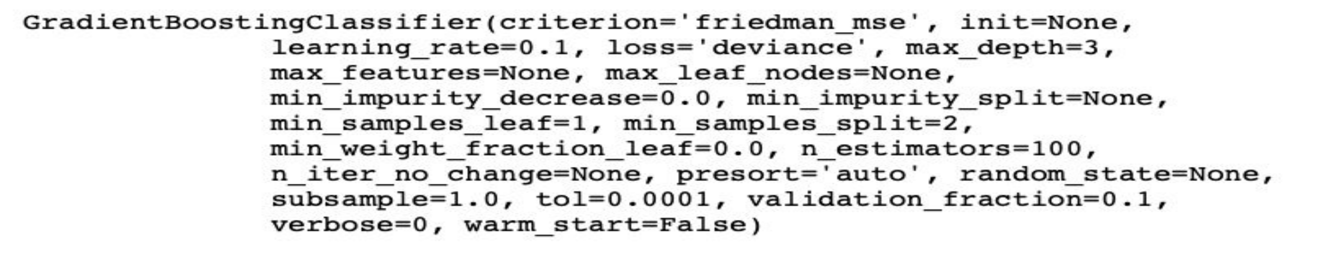
The Models



```
LogisticRegression(C=1000000000000.0, class_weight=None, dual=False,
          fit_intercept=False, intercept_scaling=1, max_iter=100,
          multi_class='warn', n_jobs=None, penalty='l2', random_state=None,
          solver='warn', tol=0.0001, verbose=0, warm_start=False)
```

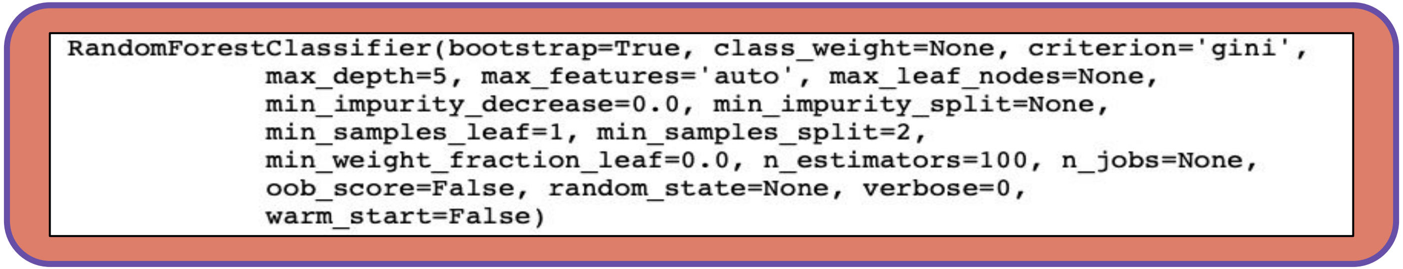


```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=10,
            splitter='best')
```

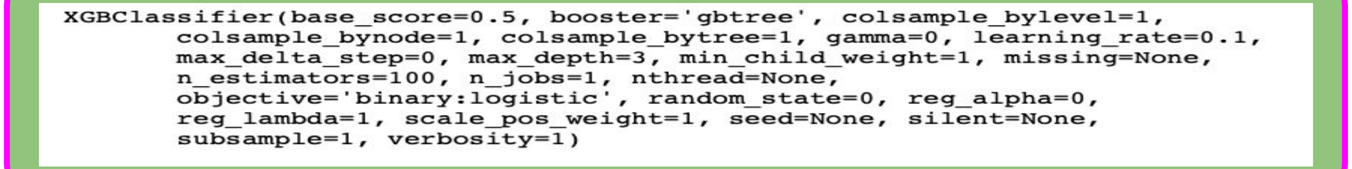


```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
              learning_rate=0.1, loss='deviance', max_depth=3,
              max_features=None, max_leaf_nodes=None,
              min_impurity_decrease=0.0, min_impurity_split=None,
              min_samples_leaf=1, min_samples_split=2,
              min_weight_fraction_leaf=0.0, n_estimators=100,
              n_iter_no_change=None, presort='auto', random_state=None,
              subsample=1.0, tol=0.0001, validation_fraction=0.1,
              verbose=0, warm_start=False)
```

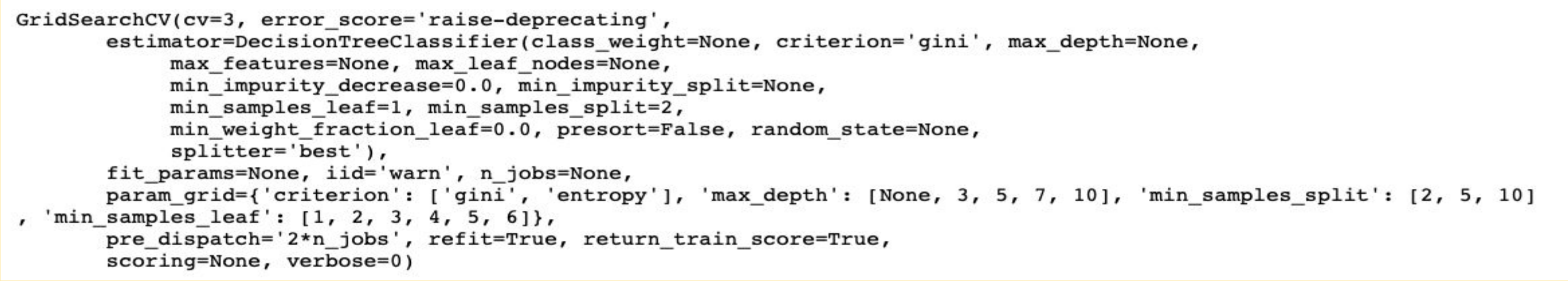
Same models but more expensive



```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=5, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=None,
                       oob_score=False, random_state=None, verbose=0,
                       warm_start=False)
```

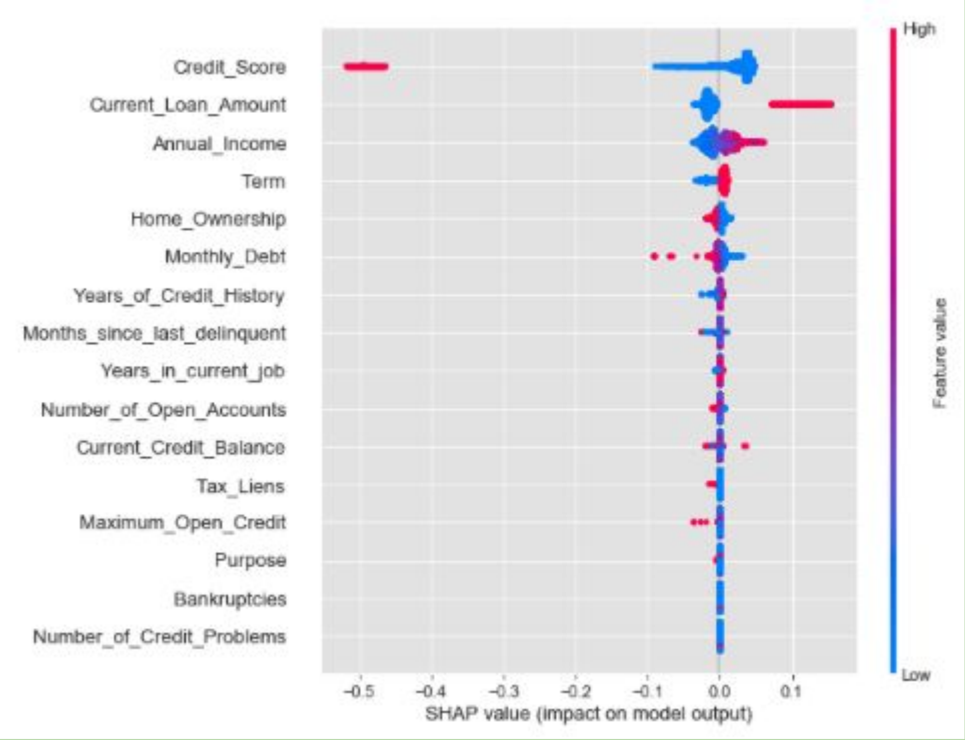


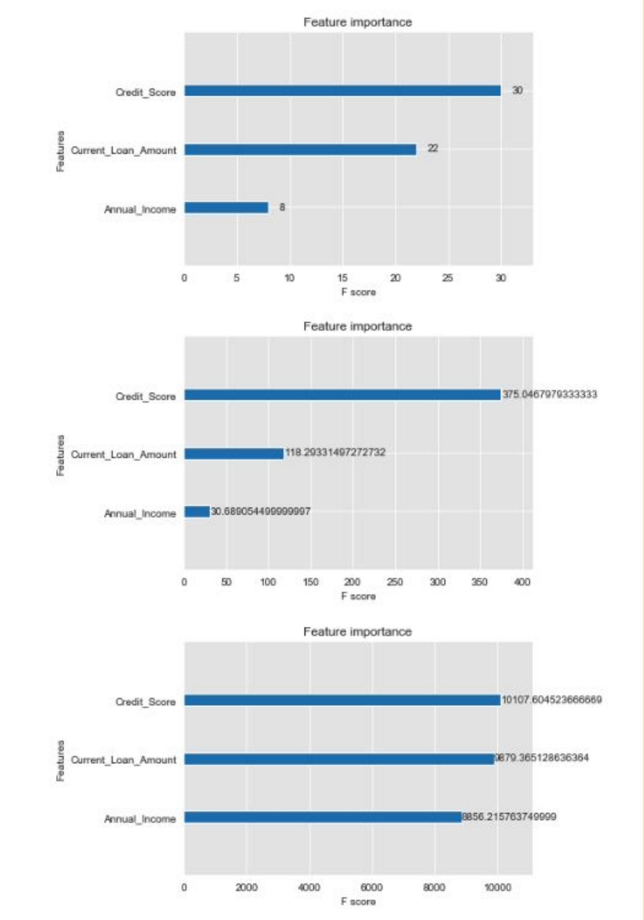
```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
       colsample_bynode=1, colsample_bytree=1, gamma=0, learning_rate=0.1,
       max_delta_step=0, max_depth=3, min_child_weight=1, missing=None,
       n_estimators=100, n_jobs=1, nthread=None,
       objective='binary:logistic', random_state=0, reg_alpha=0,
       reg_lambda=1, scale_pos_weight=1, seed=None, silent=None,
       subsample=1, verbosity=1)
```



```
GridSearchCV(cv=3, error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best'),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'criterion': ['gini', 'entropy'], 'max_depth': [None, 3, 5, 7, 10], 'min_samples_split': [2, 5, 10]
, 'min_samples_leaf': [1, 2, 3, 4, 5, 6]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring=None, verbose=0)
```

Results :


Credit_Score
Current_Loan_Amount
Annual_Income
Term
Home_Ownership
Monthly_Debt
Years_of_Credit_History
Months_since_last_delinquent
Years_in_current_job
Number_of_Open_Accounts
Current_Credit_Balance
Tax_Liens
Maximum_Open_Credit
Purpose
Bankruptcies
Number_of_Credit_Problems
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
SHAP value (impact on model output)
High
Low
Feature value


Feature importance
Features
Credit_Score
30
Current_Loan_Amount
22
Annual_Income
8
0
5
10
15
20
25
30
F score
Feature importance
Features
Credit_Score
375.0467979333333
Current_Loan_Amount
118.29331497272732
Annual_Income
30.689054499999997
0
50
100
150
200
250
300
350
400
F score
Feature importance
Features
Credit_Score
10107.604523666669
Current_Loan_Amount
9879.365128636364
Annual_Income
8856.215763749999
0
2000
4000
6000
8000
10000
F score

Further Exploration and Conclusion

More Machine Learning Models are Available

K-Nearest Neighbor, Principle Componet Analysis, Support Vector Machines

Additional Data Sets

Re-run Models on multiple data sets for accuracy and precision