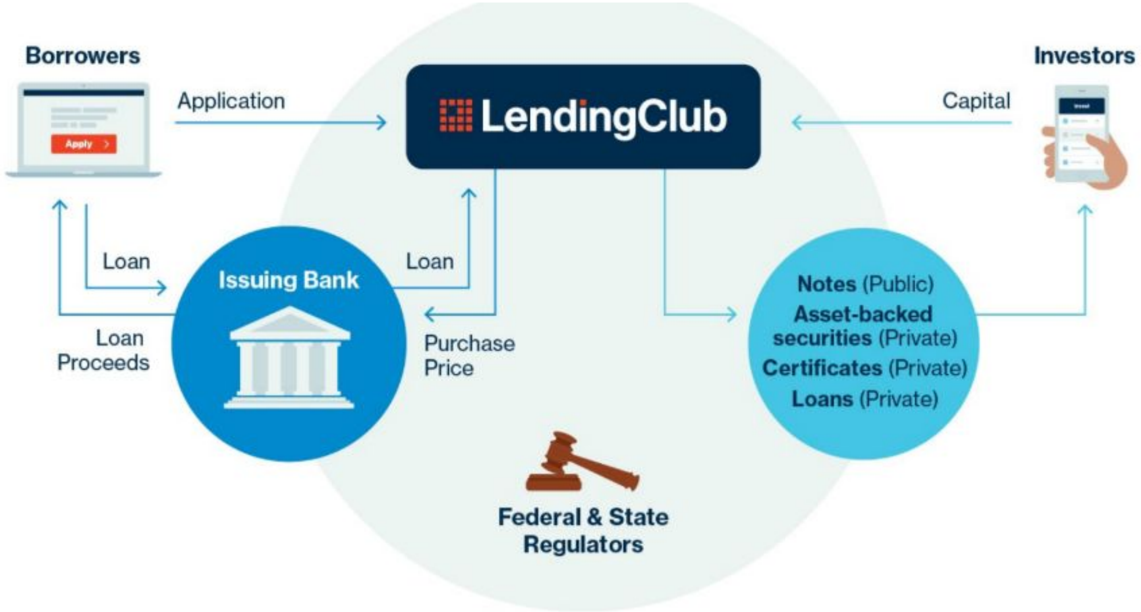
Lending Club – Predicting Loans

Gabriel Wong - 1006593018
Andrew Haddad - 1001668260
Saif Syed - 1001326647

What is Lending Club?

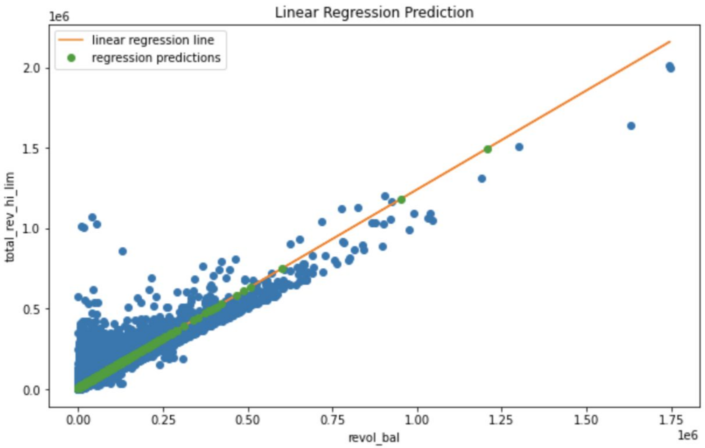


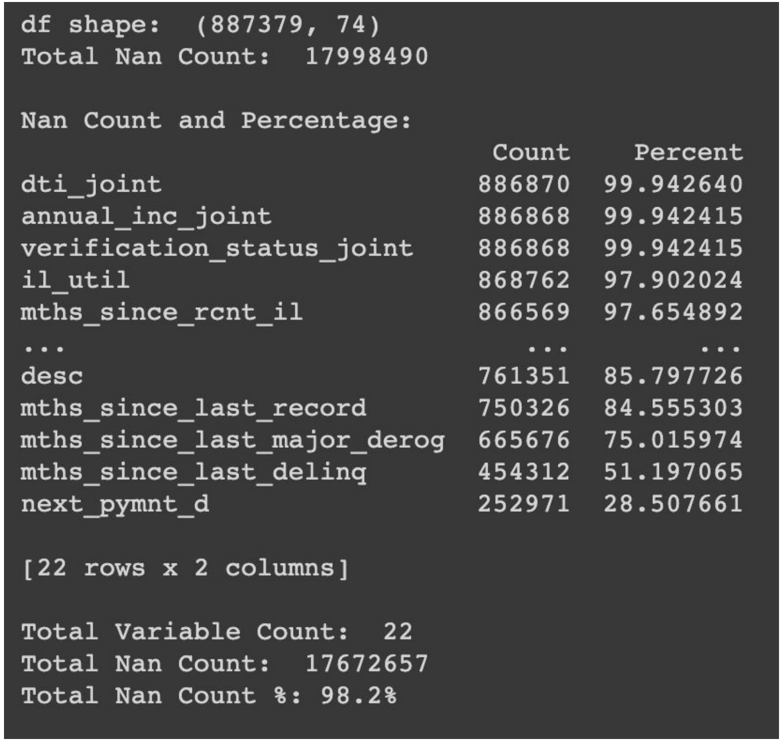
America’s Largest online lending marketplace

Objective: develop a robust and reliable supervised learning algorithm to predict whether a borrower is qualified to receive a loan or not to minimize forecasted losses

Data Cleaning and Preparation

- Removed Features with too many NaNs
- Imputed numerical features with multiple imputation with medians, zeros, or regressions
- Imputed categorical features with ‘Other”





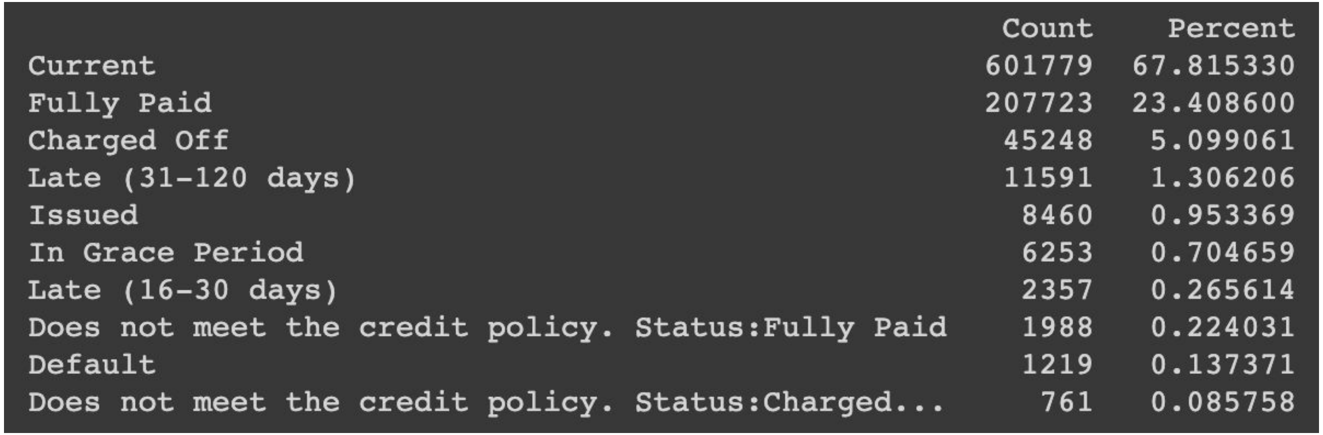
```
df shape:  (887379, 74)
Total Nan Count:  17998490

Nan Count and Percentage:
                              Count    Percent
dti_joint                     886870  99.942640
annual_inc_joint              886868  99.942415
verification_status_joint     886868  99.942415
il_util                       868762  97.902024
mths_since_rcnt_il            866569  97.654892
...                              ...        ...
desc                          761351  85.797726
mths_since_last_record        750326  84.555303
mths_since_last_major_derog   665676  75.015974
mths_since_last_delinq        454312  51.197065
next_pymnt_d                  252971  28.507661

[22 rows x 2 columns]

Total Variable Count:  22
Total Nan Count:  17672657
Total Nan Count %: 98.2%
```

Categorizing the Target Feature



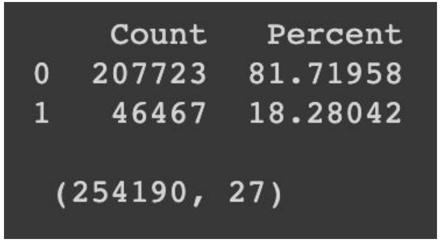
```
                                                    Count    Percent
Current                                            601779  67.815330
Fully Paid                                         207723  23.408600
Charged Off                                         45248   5.099061
Late (31-120 days)                                  11591   1.306206
Issued                                               8460   0.953369
In Grace Period                                      6253   0.704659
Late (16-30 days)                                    2357   0.265614
Does not meet the credit policy. Status:Fully Paid   1988   0.224031
Default                                              1219   0.137371
Does not meet the credit policy. Status:Charged...    761   0.085758
```

- Had to convert target feature into a context relevant feature

- Created a new target feature - ‘Good loan’ vs ‘Bad loan’

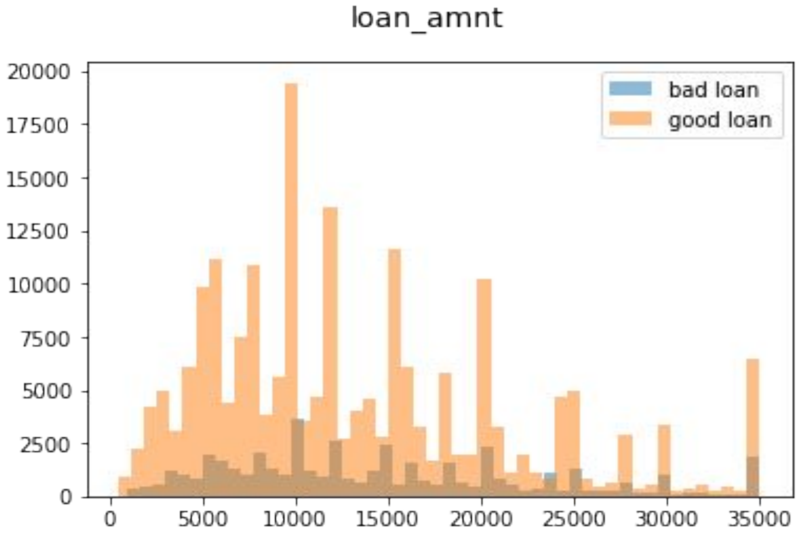
- Final input dataset count for model has ~254K entries (18% positive)

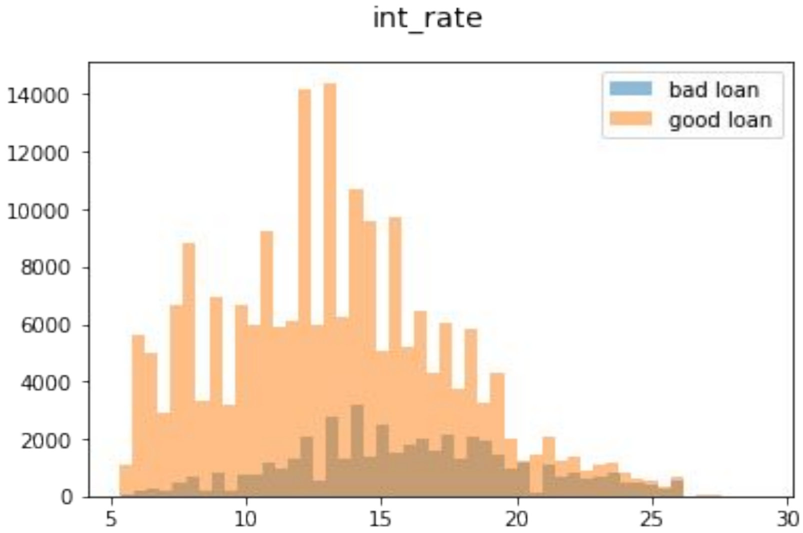
good->
bad->



```
    Count   Percent
0  207723  81.71958
1   46467  18.28042

 (254190, 27)
```

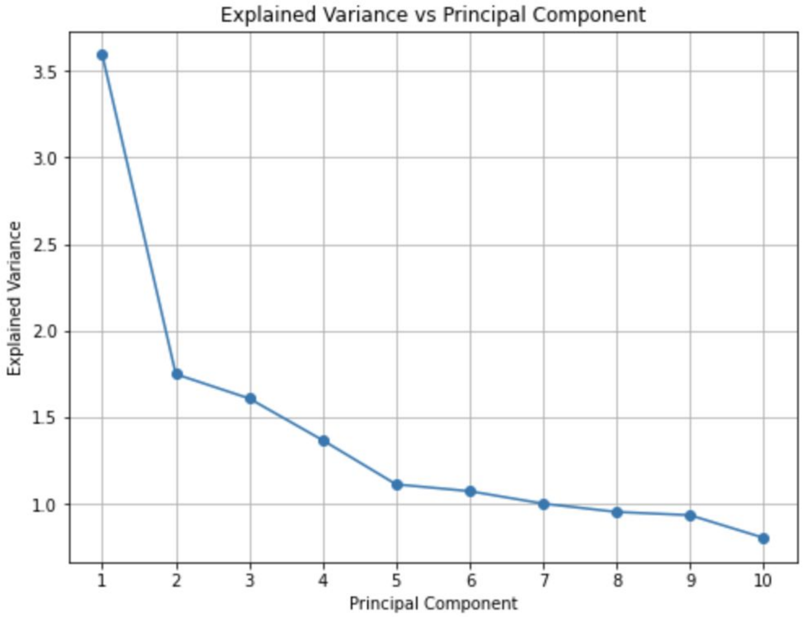


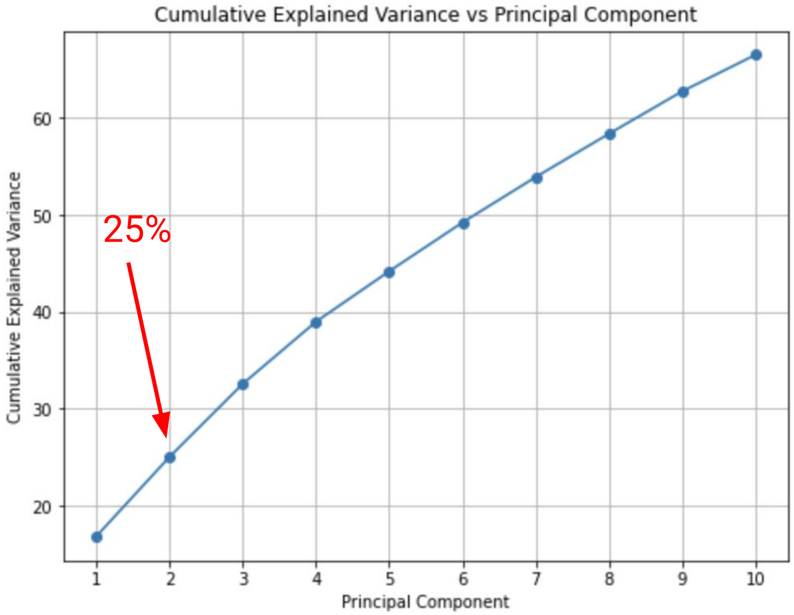


Distributions

PCA - Explained Variance

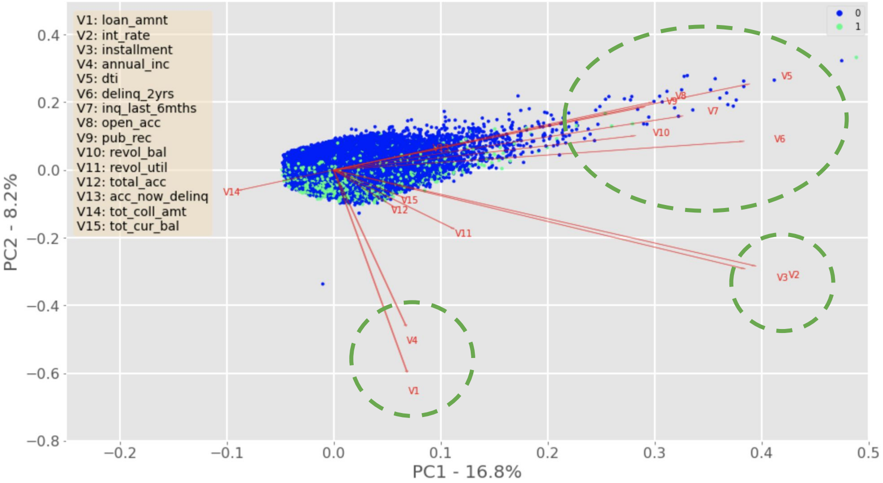
- PC1 and PC2 amount to 25% of the total explained variance.

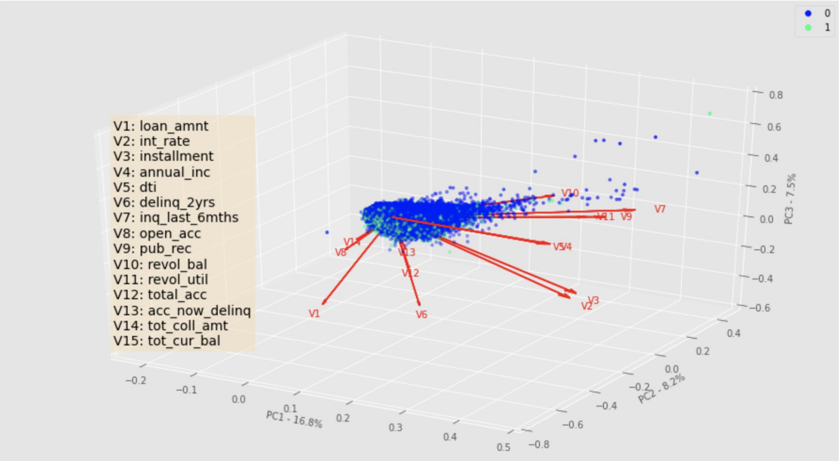




25%

PCA - Biplot & Triplot





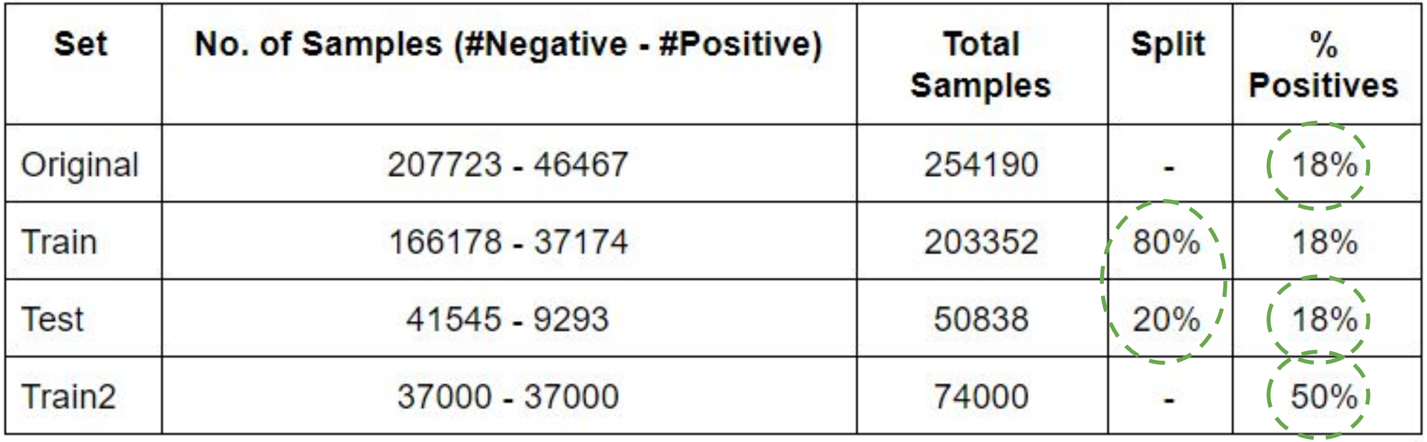
- PC1 - V5 to V10 highly correlated
- PC2 - V1 & V4 highly correlated
- V2 & V3 are somewhere in between

PCA Analysis

Features Selection:

1) Project scope - Loan pre-approval phase

2) Lasso - embedding type
   a) Train, test split & standardization
   b) Lasso - 140 features to 101 features (28% reduction)



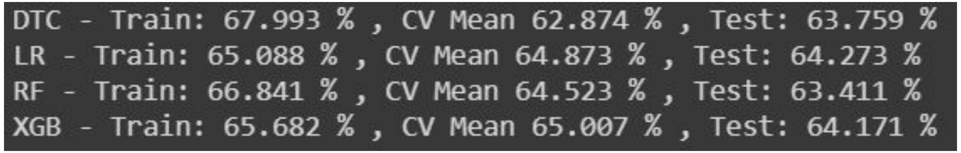
Set	No. of Samples (#Negative - #Positive)	Total Samples	Split	% Positives
Original	207723 - 46467	254190	-	18%
Train	166178 - 37174	203352	80%	18%
Test	41545 - 9293	50838	20%	18%
Train2	37000 - 37000	74000	-	50%

Feature Selection and Data Partition

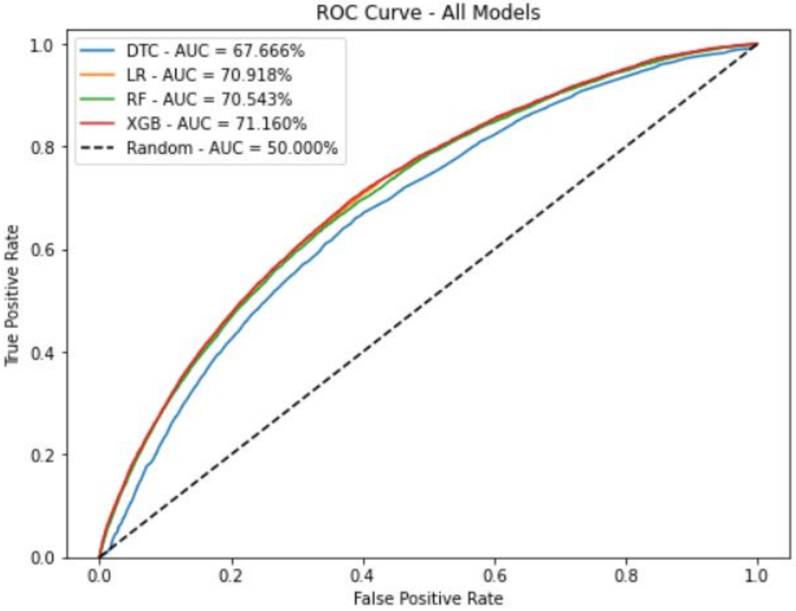
Modelling – Selection, Tuning, and Evaluation

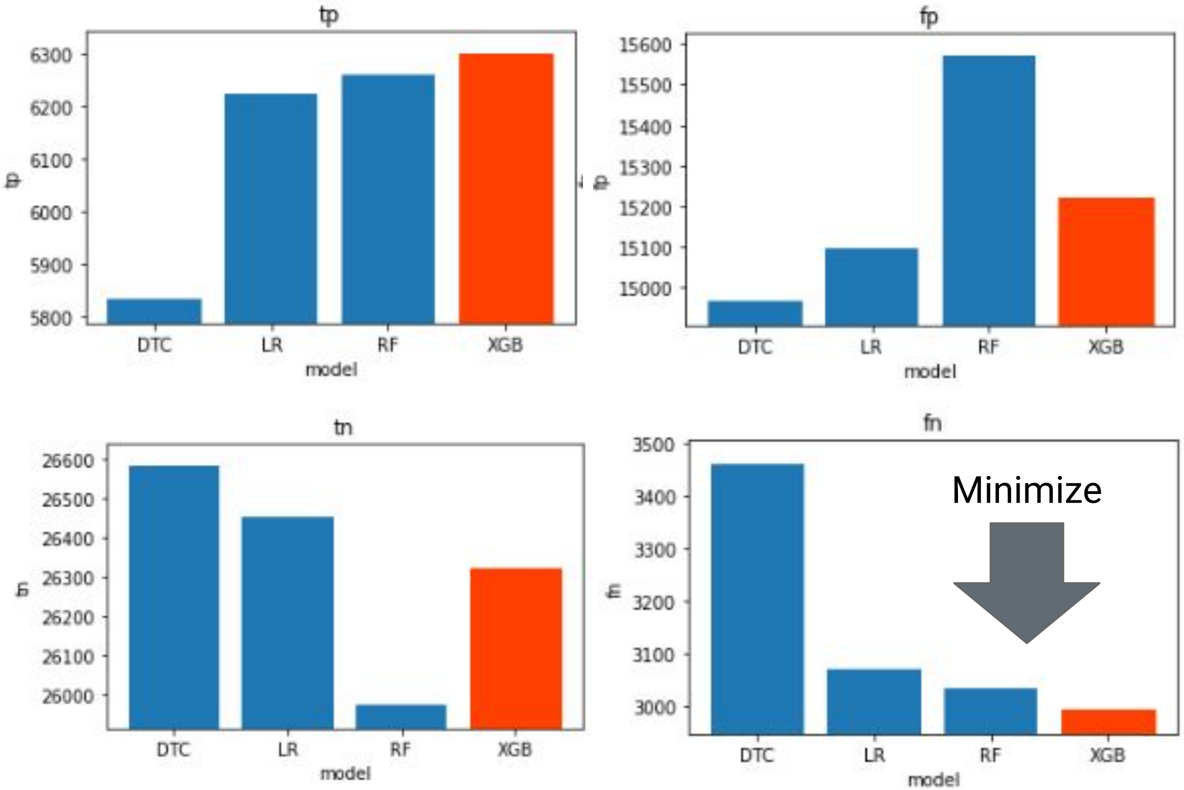
Tested four different models:

1) Decision Tree Classifier
2) Logistic Regression
3) Random Forest
4) XGBoost



```
DTC - Train: 67.993 % , CV Mean 62.874 % , Test: 63.759 %
LR - Train: 65.088 % , CV Mean 64.873 % , Test: 64.273 %
RF - Train: 66.841 % , CV Mean 64.523 % , Test: 63.411 %
XGB - Train: 65.682 % , CV Mean 65.007 % , Test: 64.171 %
```





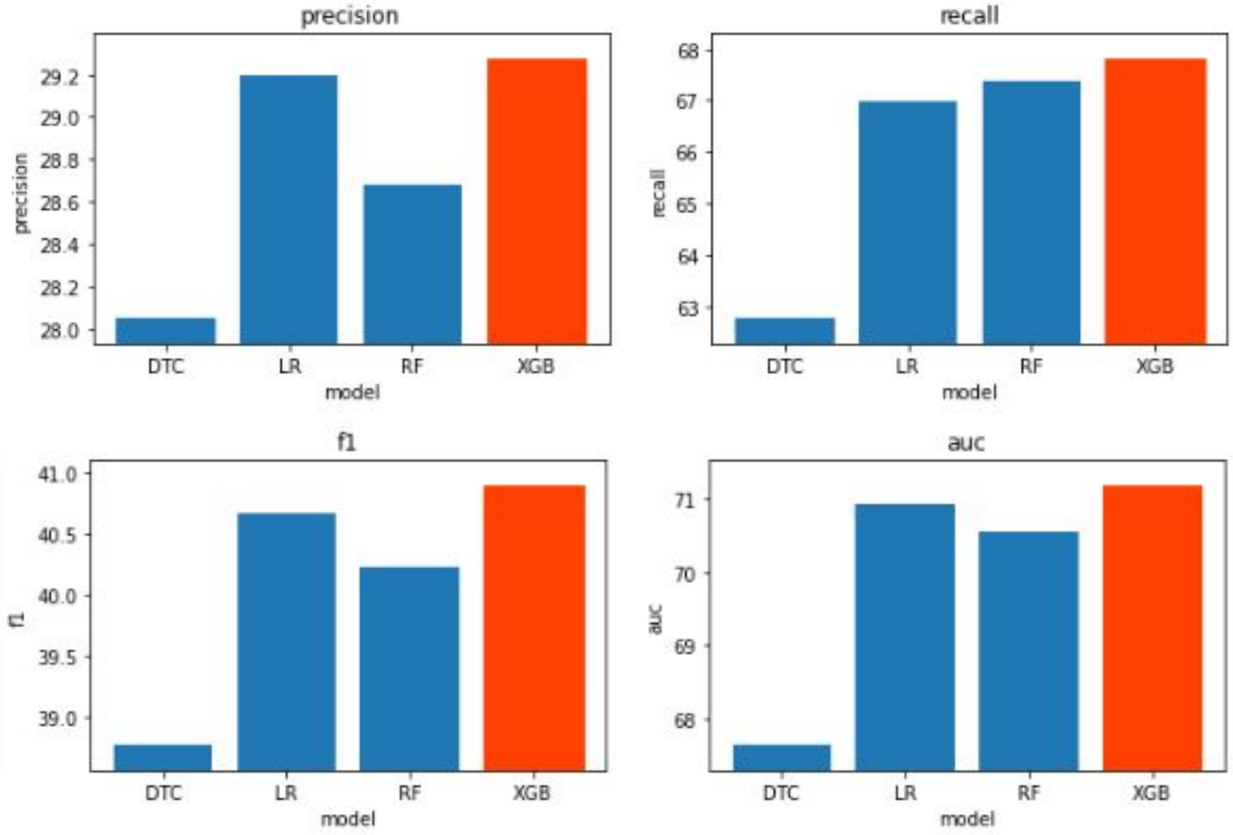
Minimize

We want the model that balances minimizing the false negative rate, while maximizing the true positive rate

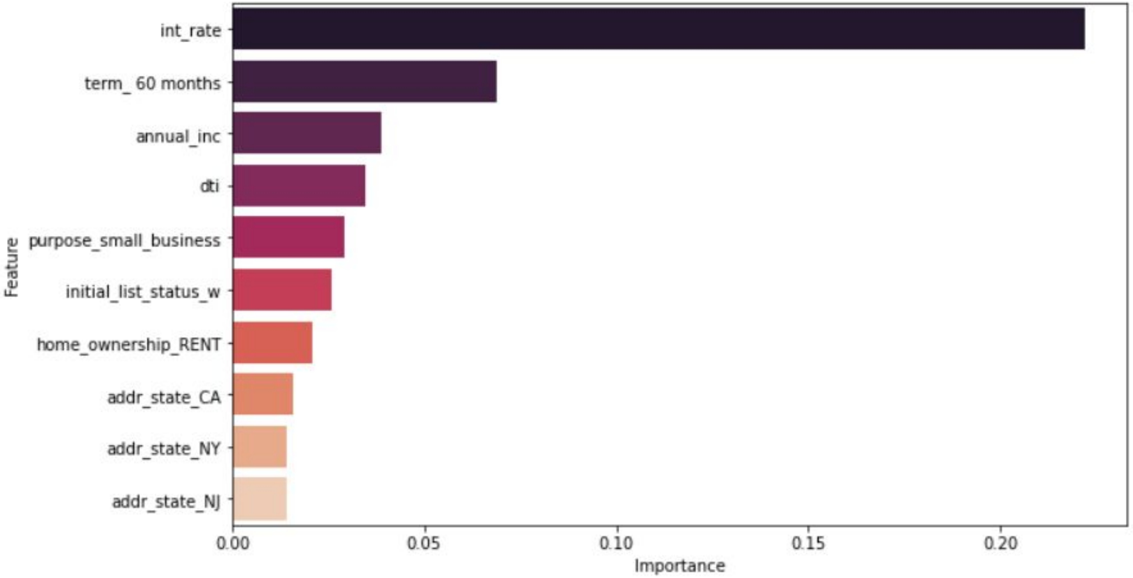
True positive tells us how correctly the model identifies bad loans

False negative tells us how many mistakes the model makes, classifying a good loan as a bad loan

Which model is best?



More Metrics for model selection



Tuning was performed using GridSearch with cross validation

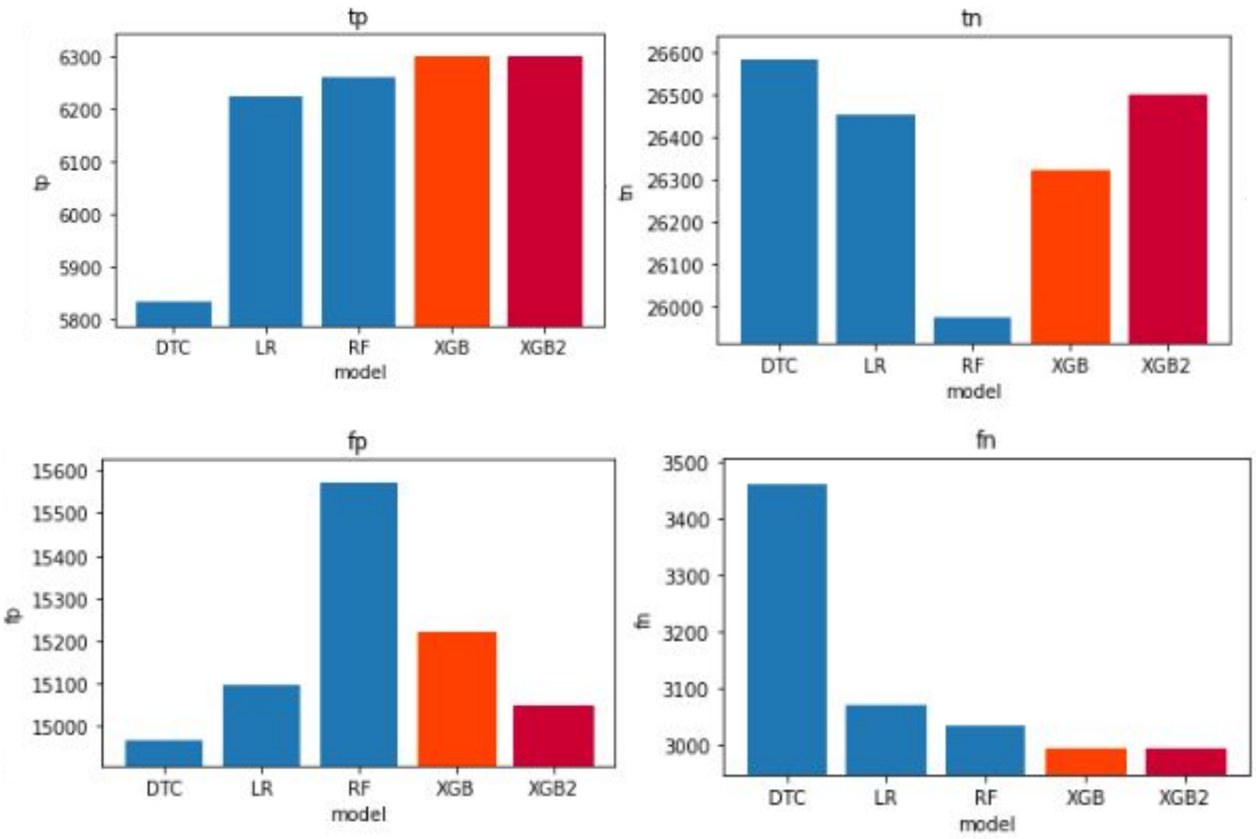
Tuning parameters:

1) Max depth
2) Min child weight
3) N estimators

Scoring criteria was ROC-AUC, rather than accuracy

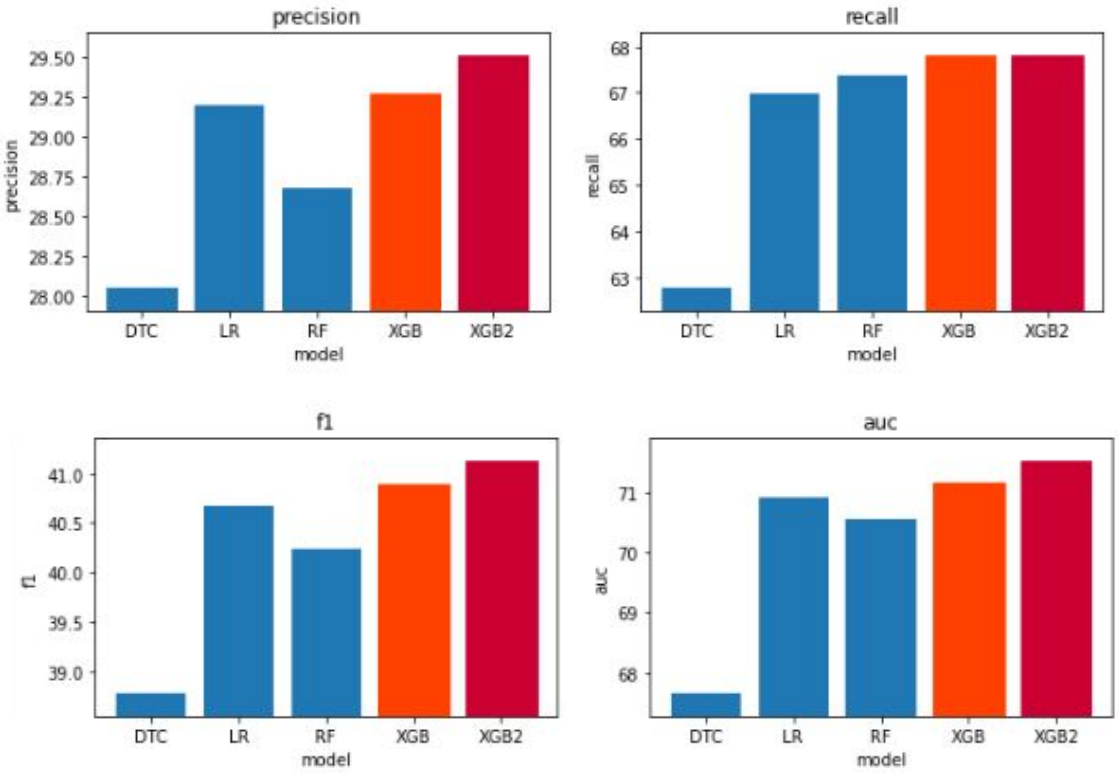
Interest rate, loan term of 60 months, and annual income found to be the most important features for the XGBoost model

Feature Importance and Model Tuning



Tuning Results in a model with a lower False positive, and higher true negative

Evaluation



precision

recall

f1

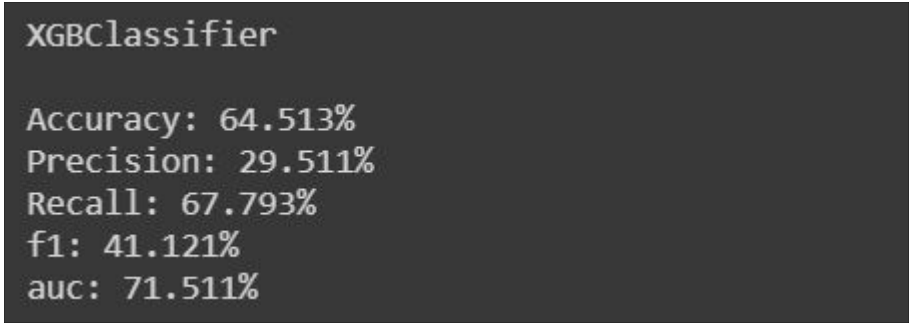
auc

More Evaluation metrics

Conclusion

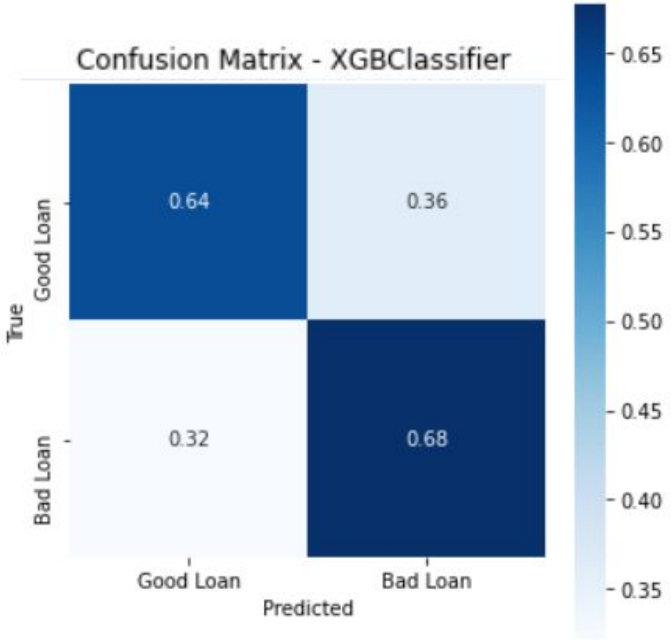
We developed an XGB classifier model to help determine which loans are likely to provide returns for the company

Next step: Deploy and Calculate Lift



```
XGBClassifier

Accuracy: 64.513%
Precision: 29.511%
Recall: 67.793%
f1: 41.121%
auc: 71.511%
```


Confusion Matrix - XGBClassifier
True
Good Loan
Bad Loan
Predicted
Good Loan
Bad Loan