





DANIEL MARQUEZ

DATA SCIENCE CAPSTONE PROJECT

SPRINGBOARD 2021 COHORT

LOAN DEFAULT PREDICTIVE MODEL

THE PROBLEM

- ▶ The purpose of this data science project is to have a better selection process for individuals who apply for a loan. overall, there are 4 main outcomes for a finance company when loaning a customer money
1. Deny the applicant a loan when they could have paid 
 2. Approve the applicant and they default. 
 3. Approve the applicant, applicant pays back loan on time in full 
 4. Deny applicant, applicant would have defaulted anyway 

STAKE HOLDERS

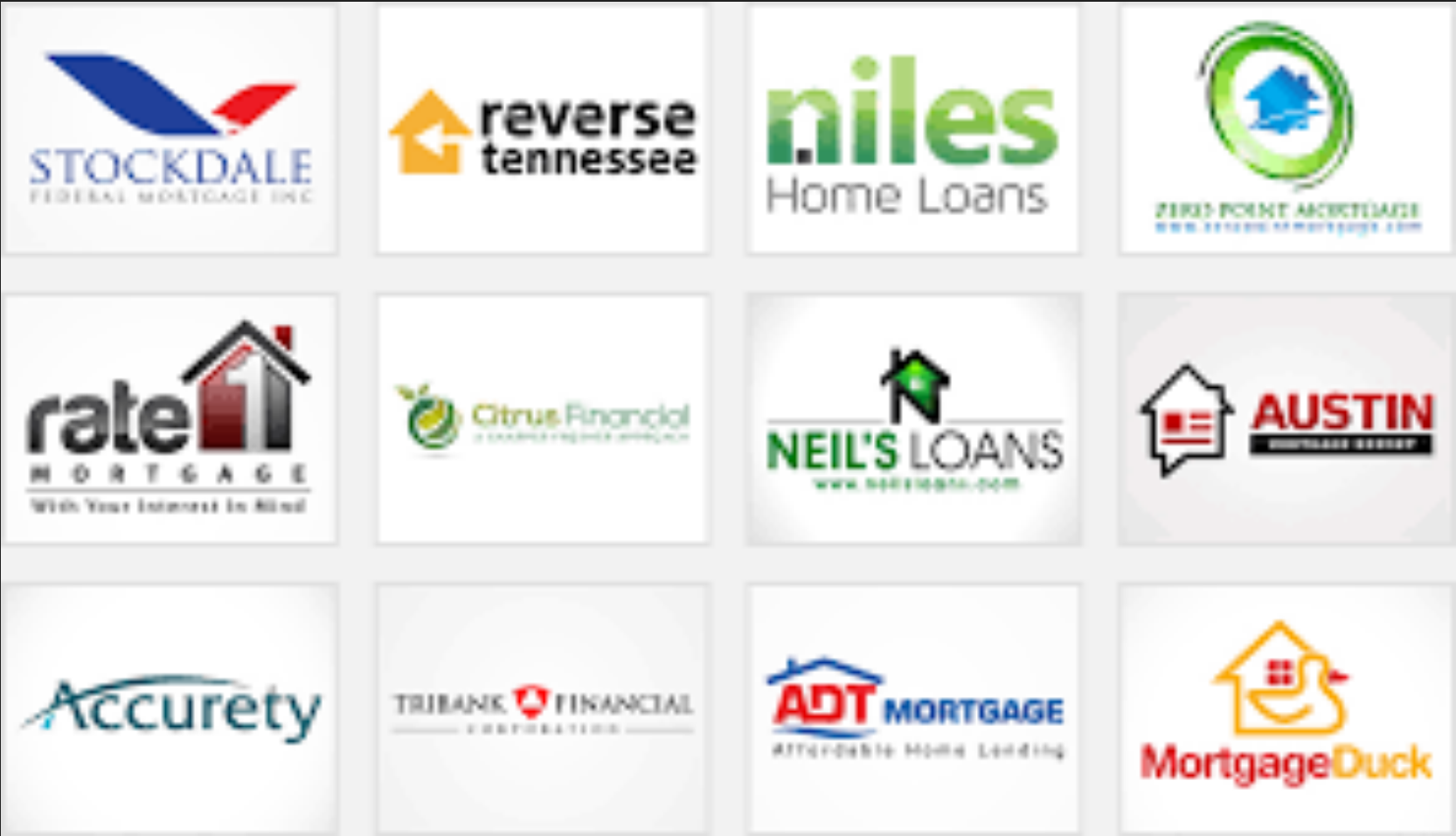
CREDIT CARD COMPANIES



FINANCIAL SERVICE'S



MORTGAGE COMPANIES



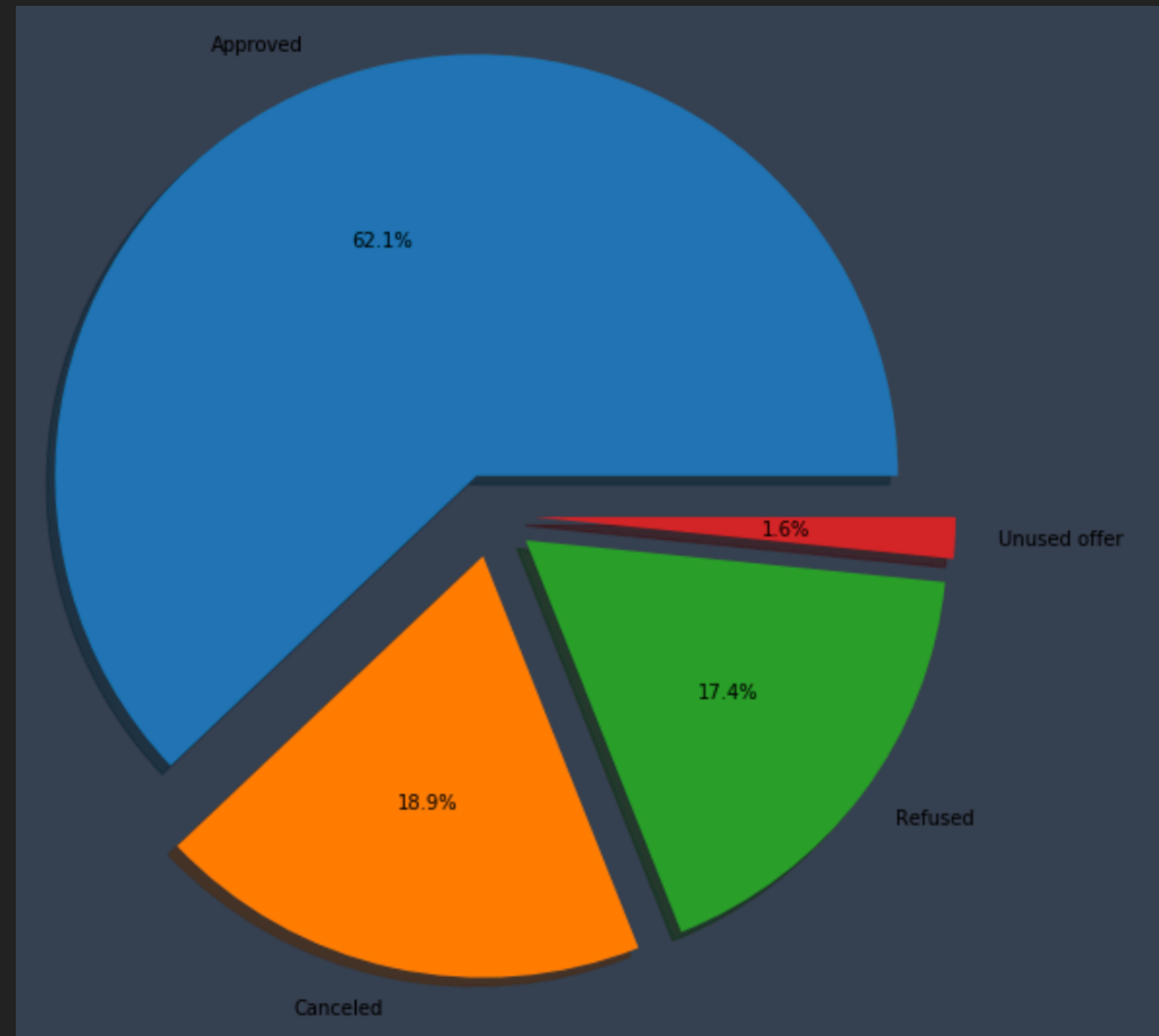
INFORMATION ABOUT THE DATA

The data set was obtained from kaggle and is a real world data set and contains two different sets of data, one of current applications and the other is of past applicants. Some manipulation and merging is needed in order to get the data to be informative. a link to the kaggle data is shown below

<https://www.kaggle.com/gauravduttakiit/loan-defaulter>

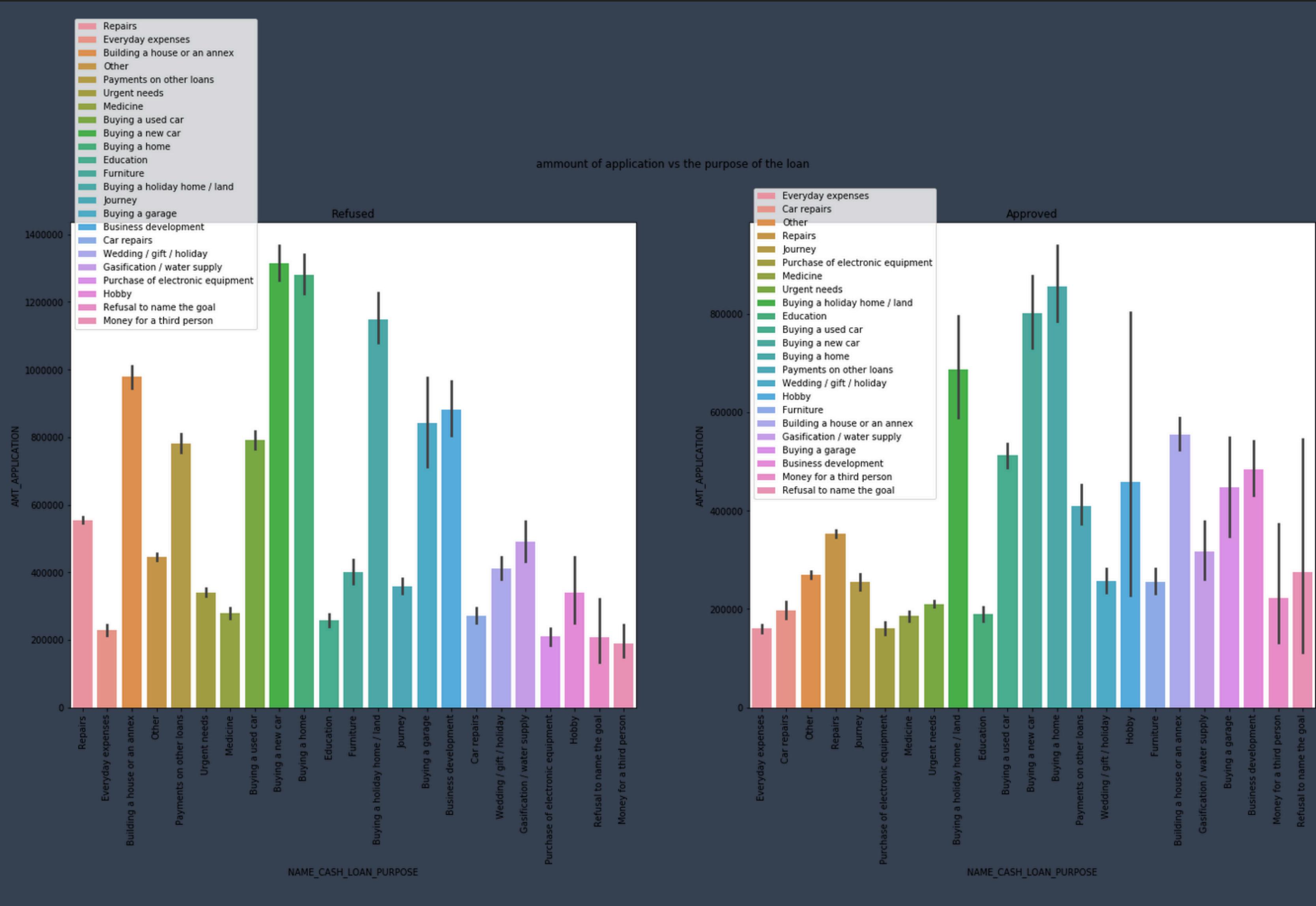
DATA EXPLORATION

- ▶ Most of the applications are approved



DATA EXPLORATION

- Approved and refused loans categorized by category and loan amount



MACHINE LEARNING MODELING

CLASSIFICATION MODELS USED

1. RANDOM FOREST

2. XGBOOST

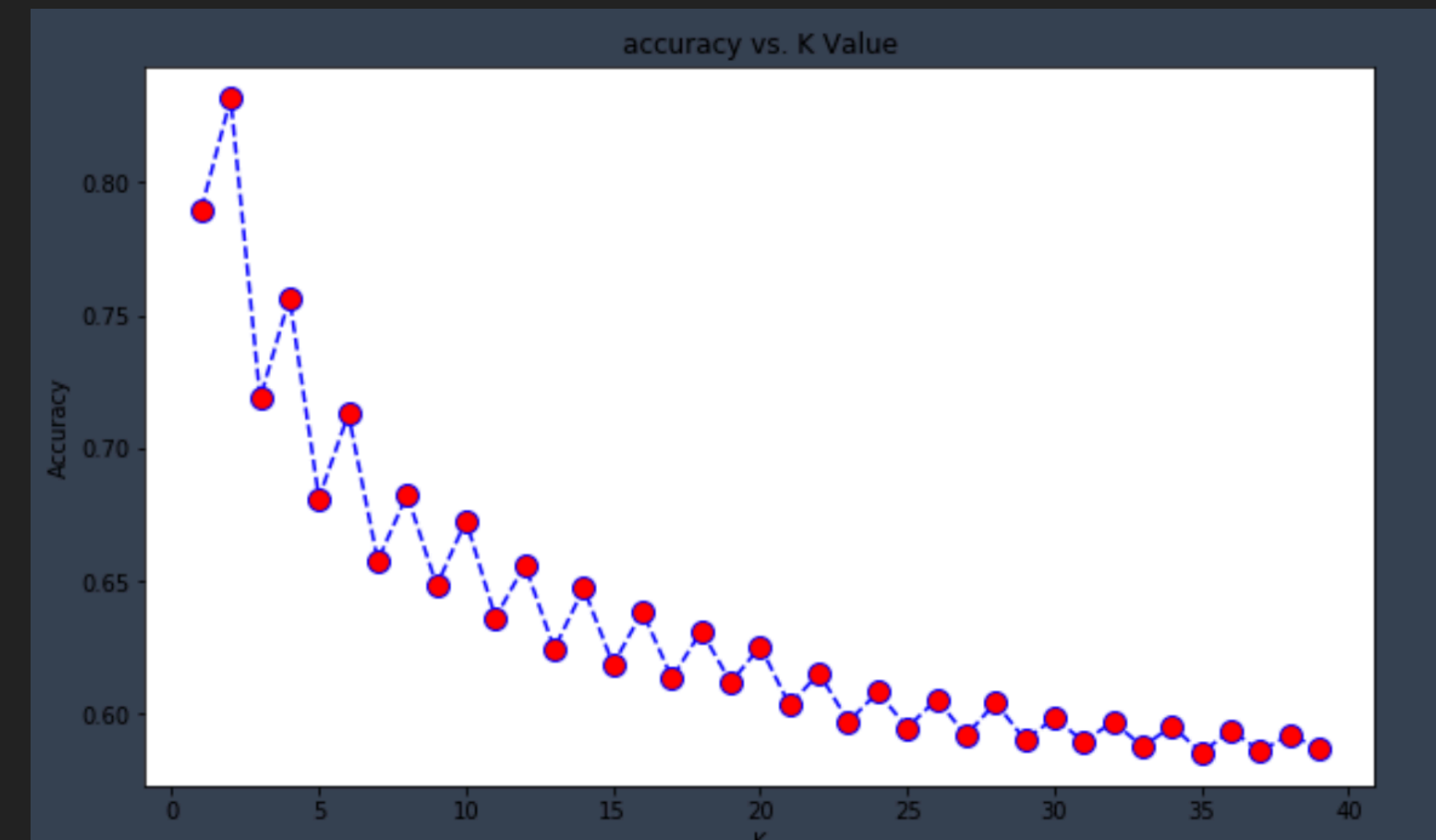
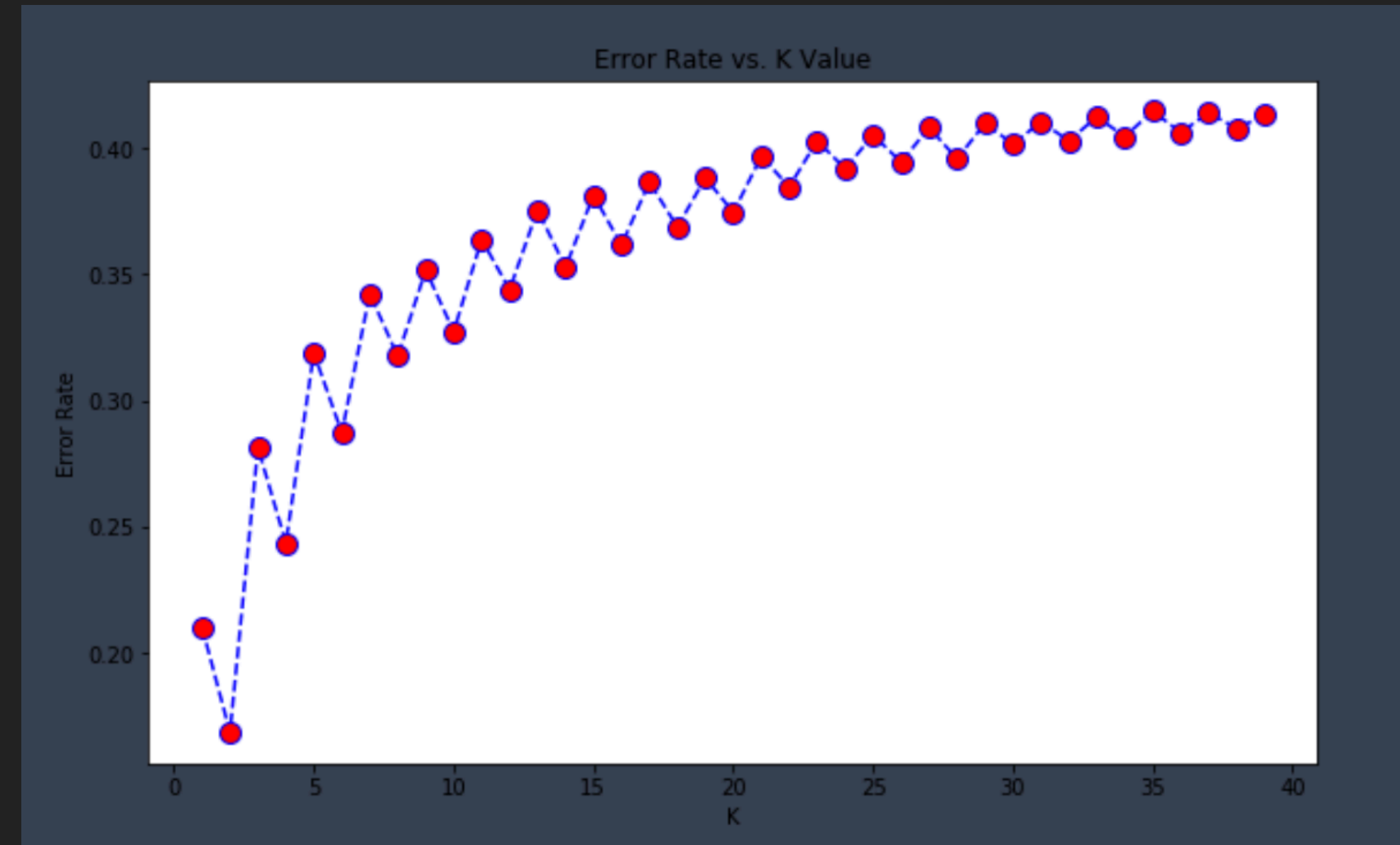
3. KNN

- ▶ TRAIN WITH DATA BALANCED DATA SET, IMPLEMENTED SMOTE (Synthetic Minority Oversampling Technique)
- ▶ 20% OF DATA WAS SET OFF TO THE SIDE AND LEFT UN BALANCED, OF THE 80% DATA LEFT OVER, 70% OF THAT WAS FOR TESTING AFTER APPLYING SMOTE WHILE THE REMAINING WAS USED AS A FIRST ROUND OF TESTING
- ▶ COMPARE END RESULTS WITH SUBSET OF UNBALANCED DATA AND MOVE FORWARD WITH THE BEST PERFORMING MODEL
- ▶ KNN WITH $K=11$ PERFORMED THE BEST

KNN CLASSIFIER

GRID SEARCH WAS FIRST IMPLEMENTED AND RETURNED A K-VALUE OF 11 WOULD BE THE OPTIMAL VALUE TO MOVE FORWARD WITH. JUST TO VERIFY THE ERROR RATE V.S. K-VALUE WAS PLOTTED. AS WE CAN SEE, AS K INCREASES THE ERROR RATE TENDS TO RISE AS WELL AS SHOWN BELOW.

MOVED FORWARD WITH $K = 11$



MODELING PERFORMANCE COMPARISONS

MODEL PERFORMANCE ON THE 20% BALANCED DATA(SMOTE)



KNN

TEST DATA ACCURACY.....63.89 %

TEST DATA SENSITIVITY.....98.92 %

TEST DATA SPECIFICITY.....31.52 %



XGBOOST CLASSIFIER

TEST DATA ACCURACY.....95.8 %

TEST DATA SENSITIVITY.....91.25 %

TEST DATA SPECIFICITY.....100.0 %



RANDOM FOREST CLASSIFIER

TEST DATA ACCURACY.....95.8 %

TEST DATA SENSITIVITY.....91.25 %

TEST DATA SPECIFICITY.....100.0 %

MODEL PERFORMANCE ON THE 20% UNBALANCED DATA



KNN

TEST DATA ACCURACY.....37.36 %

TEST DATA SENSITIVITY.....71.11 %

TEST DATA SPECIFICITY.....33.75 %



XGBOOST CLASSIFIER

TEST DATA ACCURACY.....92.07 %

TEST DATA SENSITIVITY.....7.0 %

TEST DATA SPECIFICITY.....98.62 %



RANDOM FOREST CLASSIFIER

TEST DATA ACCURACY.....92.79 %

TEST DATA SENSITIVITY.....2.0 %

TEST DATA SPECIFICITY.....99.77 %

IMPROVEMENTS

- ▶ THIS DATA IS A SUBSET OF A DIFFERENT DATA SET USED IN A COMPETITION 3 YEARS AGO WHERE THE WINNER WINS UP TO \$70,000
[HTTPS://WWW.KAGGLE.COM/C/HOME-CREDIT-DEFAULT-RISK](https://www.kaggle.com/c/home-credit-default-risk)
I BELIEVE USING THIS SET OF COMPETITION DATA MAY LEAD TO SOME INTERESTING RESULTS GIVEN HOW MUCH MORE DATA THERE IS TO FACILITATE.
- ▶ ENGINEER MORE FEATURES INDICATIVE OF POTENTIAL DEFAULTERS IN ORDER TO SPOT PATTERS WITHIN THE APPLICANTS
- ▶ EXPLORE THE LINKED DATA IN EFFORTS OF OBTAINING MORE DATA ABOUT THE APPLICANTS
- ▶ ATTEMPT OTHER METHODS THAT COMBAT THE UNBALANCED DATA SET SUCH AS RUS(RANDOM UNDER-SAMPLING), OR SKLEARN'S CLASS_WEIGHT(='BALANCED') TECHNIQUE
- ▶ ATTEMPT OTHER CLASSIFICATION MODELS SUCH AS GAUSSIAN NAIVE BAYES OR EXTREMELY RANDOMIZED TREES

CONCLUSION

- ▶ OF THE THREE SUPERVISED CLASSIFICATION MODELS, KNN WITH $K = 11$ PERFORMED THE BEST HOWEVER, THERE IS STILL ROOM FOR IMPROVEMENT
- ▶ MODELS WERE EVALUATED BASED ON THE 20% UNBALANCED DATA INITIALLY SET ASIDE