Loan Status Analysis

Max Groves

Introduction

I used a set of Lending Club Data for my Project

My target variable throughout the Project was the Current Loan Status

I used three different models:

1. KNN
2. Logistic Regression
3. Decision Tree

Changing Loan Status

```
In [15]: loans.loc[:,'loan_status'].value_counts()

Out[15]: Current                                         601779
         Fully Paid                                      207723
         Charged Off                                      45248
         Late (31-120 days)                               11591
         Issued                                            8460
         In Grace Period                                   6253
         Late (16-30 days)                                 2357
         Does not meet the credit policy. Status:Fully Paid    1988
         Default                                           1219
         Does not meet the credit policy. Status:Charged Off    761
         Name: loan_status, dtype: int64
```

I wanted to simplify the target variable so I took the loan statuses on the left and mapped them to a Bad Status indicator

Logistic Regression

With 2 possible end states I decided to use a logistic regression model as my first go around

My Feature Columns were loan amount, interest rate, size of payment, dummy variables for housing status, term, annual income and number of open accounts

I used a Train Test Split to train and fit my model

The score ended up being 93.06% - This initially looked pretty good BUT after looking at the actual predictions it seemed the model was just predicting Good loan status every time

The outcome of this fit was the NULL model.

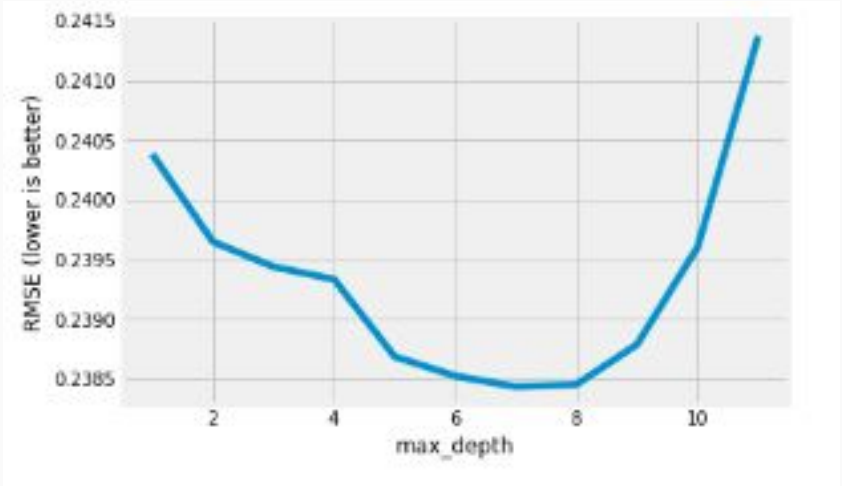
I expanded the model to try and fit on ALL the data set to see if there was some type of bias in test set selection but that model returned the same result!

Decision Tree Model

I used the same feature columns as I did for the Logistic Regression Model

In order to select a the most accurate model I built a loop and recorded the RMSE values of each depth (graph on the right)

I found my best model at a Max depth of 7, and fitted a model to that



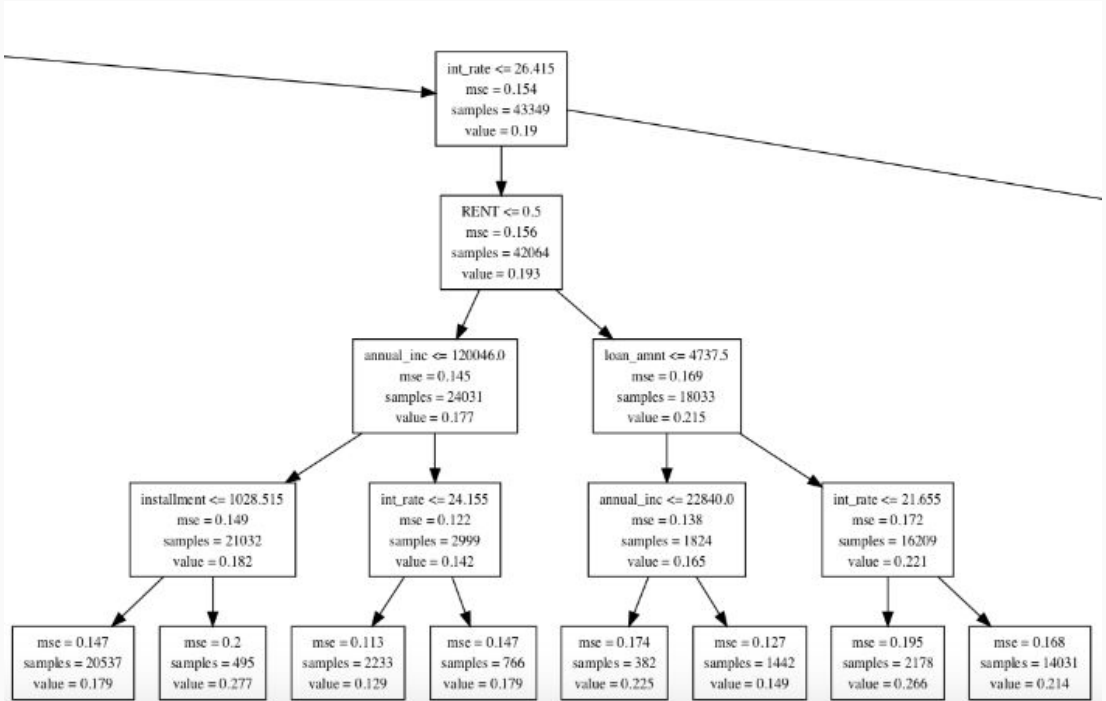
Decision Tree Results

Once fitting my model I wanted to see which factors were most important so I grabbed the feature importances

My model output was that Interest Rate was the most important factor, with annual income 60 month term, installment (payment amount), loan amount, the Rent dummy and open accounts also having some importance. This mostly seems like a reasonable result

	feature	importance
1	int_rate	0.877441
8	annual_inc	0.041783
7	60 months	0.030006
2	installment	0.028336
0	loan_amnt	0.010927
6	RENT	0.007850
9	open_acc	0.003657
3	NONE	0.000000
4	OTHER	0.000000
5	OWN	0.000000

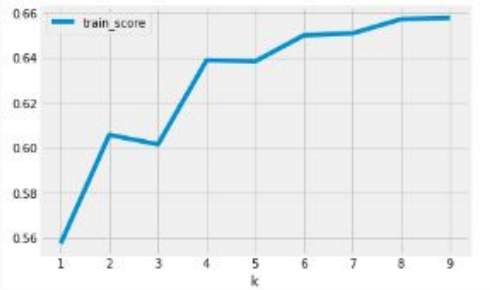
Decision Tree Picture (Just a portion)



Side Note: KNN

```
In [15]: loans.loc[:,'loan_status'].value_counts()

Out[15]: Current                                        601779
         Fully Paid                                     207723
         Charged Off                                     45248
         Late (31-120 days)                              11591
         Issued                                           8460
         In Grace Period                                  6253
         Late (16-30 days)                                2357
         Does not meet the credit policy. Status:Fully Paid    1988
         Default                                          1219
         Does not meet the credit policy. Status:Charged Off    761
         Name: loan_status, dtype: int64
```



I tried using a KNN model on the original loan status as KNN can also account for multiple end states (All the states are in the table on the left)

My best model (with a train test split) ended up being a 1 nearest neighbor (left, bottom)

Summary

After looking at the outcome of my 2 main models it seemed like the Decision Tree was easily the best performer

Having Interest Rate be the most important factor makes sense as people who have higher interest rates are usually riskier customers, and will have to pay back more interest for similar purchases than others

Possible next steps for this analysis would have been looking at collections of the users with high interest rates, and possibly running a test to see if offering these users different terms would lead to less of them getting to bad status

Thanks!