Capstone

by Jeremy Hart

Problem Statement

- Can you tell whether someone fully paid of their loan?
- Are people with different employment roles more inclined to fully pay of their loan?

https://upload.wikimedia.org/wikipedia/commons/thumb/e/e2/Wells_Fargo_Logo_%282020%29.svg/150px-Wells_Fargo_Logo_%282020%29.svg.png

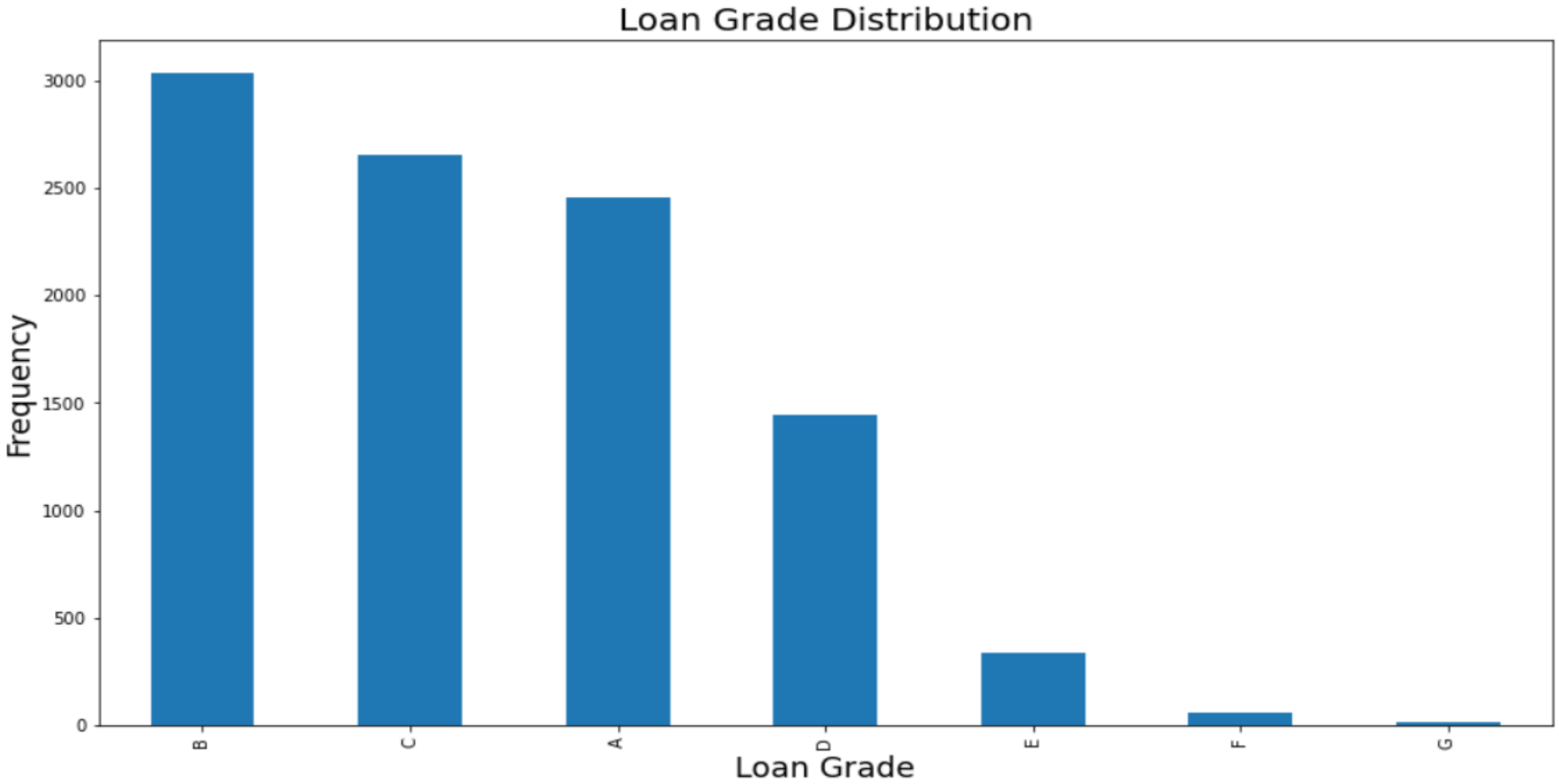
Banks and Loans

- Wells Fargo is a nationwide bank and has locations in 37 states
- Not all banks offer personal loans
- Wells Fargo offers 3 different personal loan plans
- Late payments and Defaulting is problematic

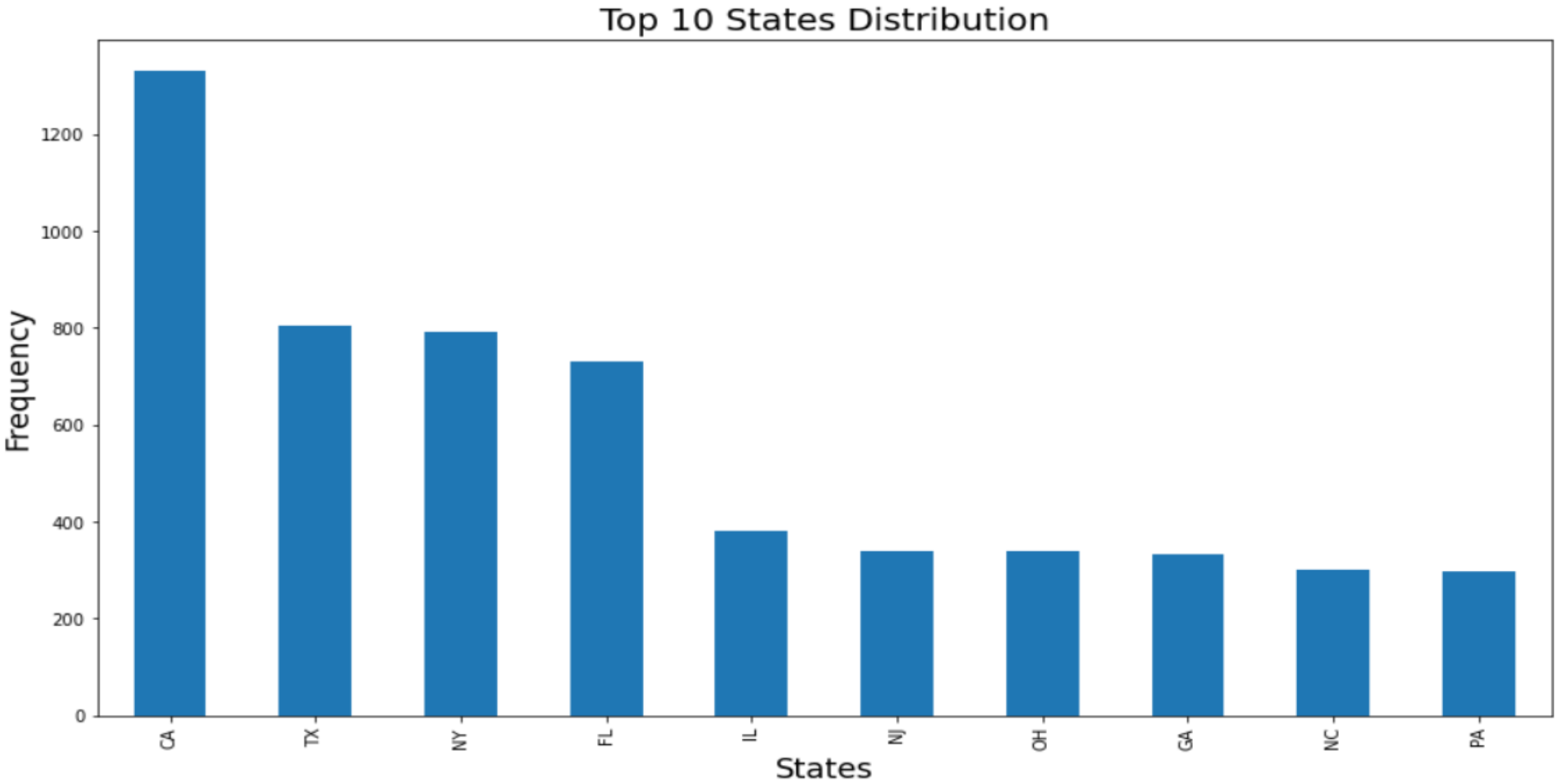
Data Cleaning

- Not too many missing values
- Feature engineering was necessary
- Made additional columns to check for correlation
- Found out the fully paid column was highly correlated with some features

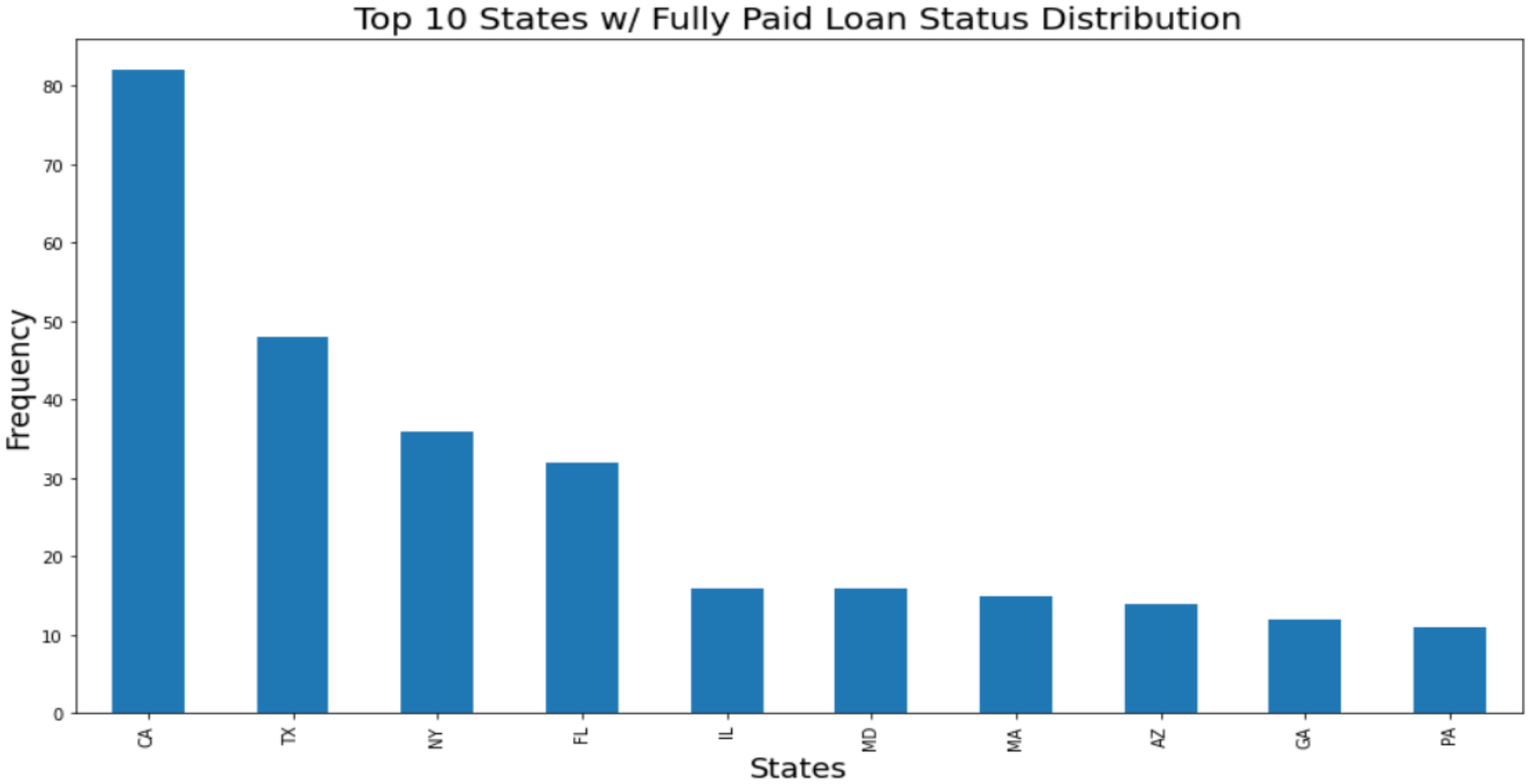
Loan Grade


Loan Grade Distribution
Frequency
3000
2500
2000
1500
1000
500
0
B
C
A
D
E
F
G
Loan Grade

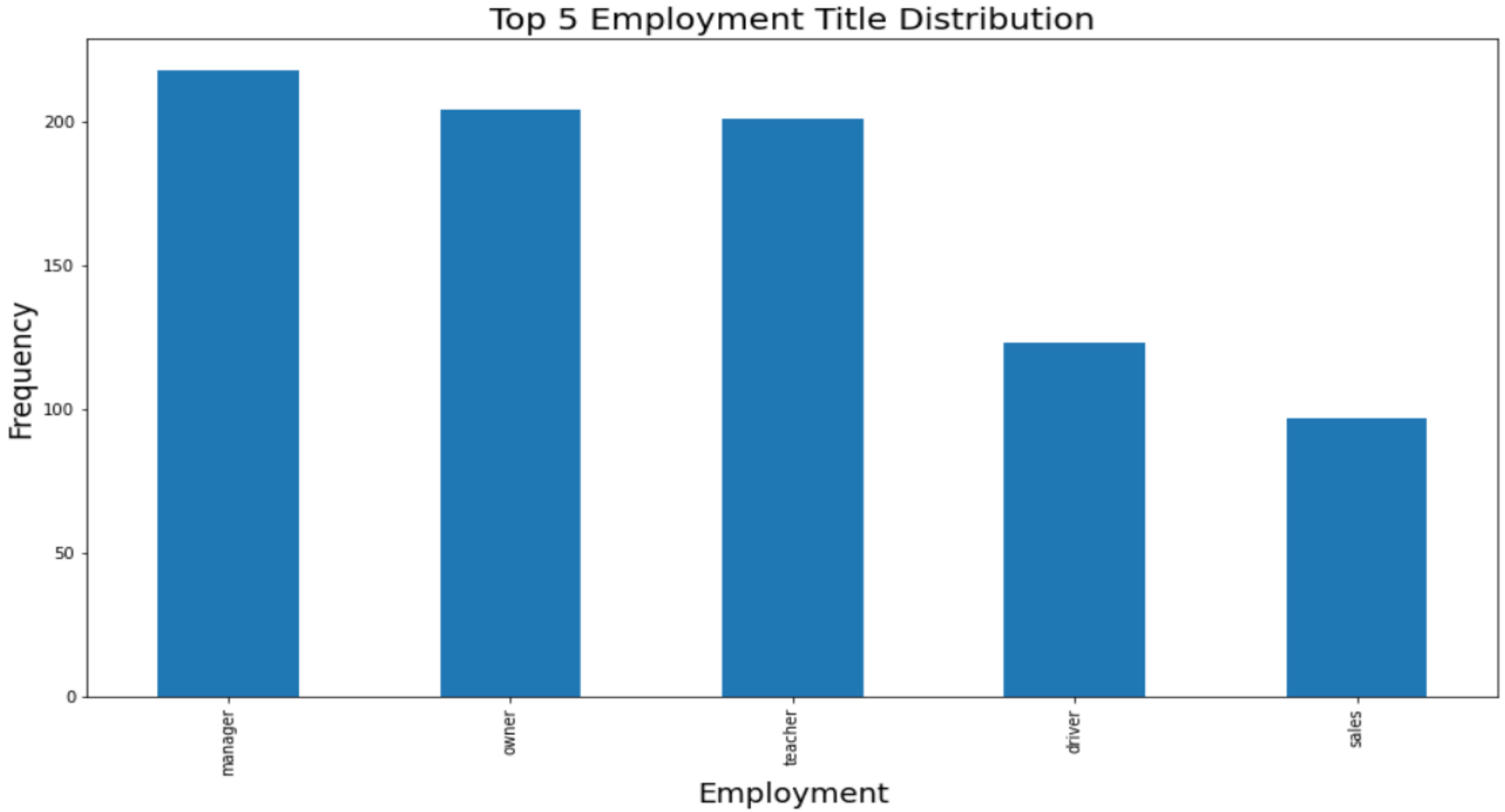
Loan State Distribution



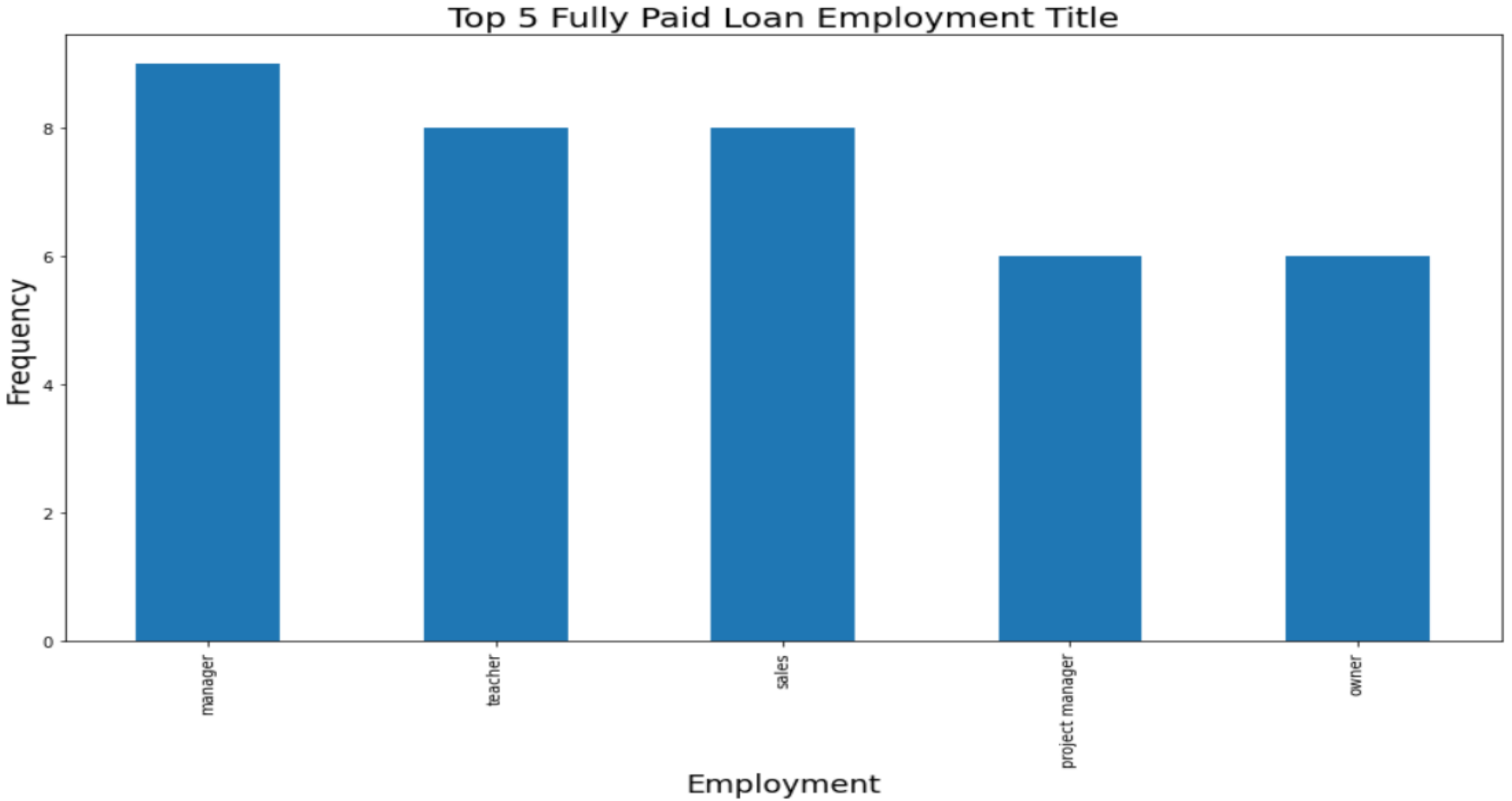
Loan States Distribution(Fully Paid)



Employment Title Distribution



Employment Title Distribution(Fully Paid)



Model Preparation

- Imbalanced data
- RandomUnderSampling Strategy
- Before resampling model was not picking up on both classes

```
0    9553
1     447
```

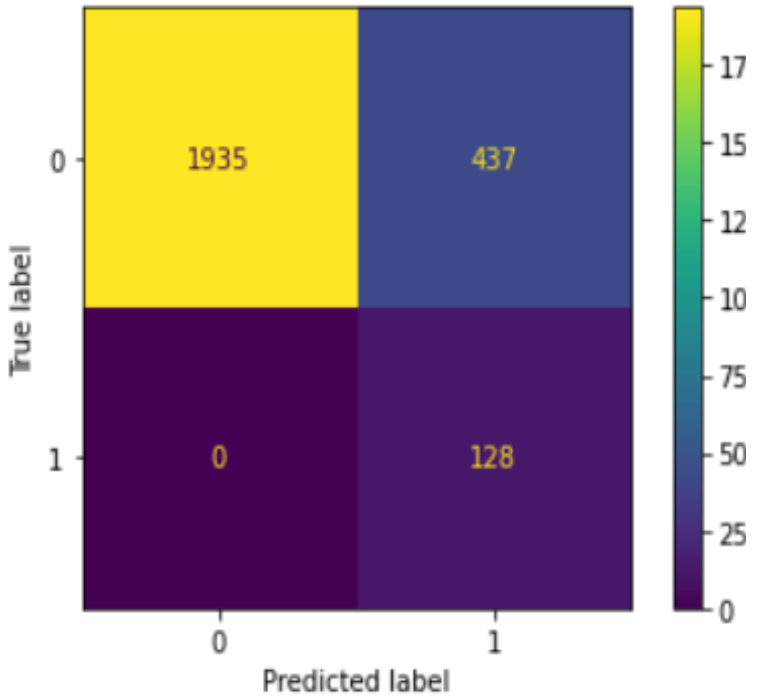
Models(w/resampled strategy)

- K-nearest neighbor, Adaboost w/Decision Tree, and Multinomial Naïve Bayes models all were resampled
- The models fit poorly on the training data
- Only predicted for the Current loan status class

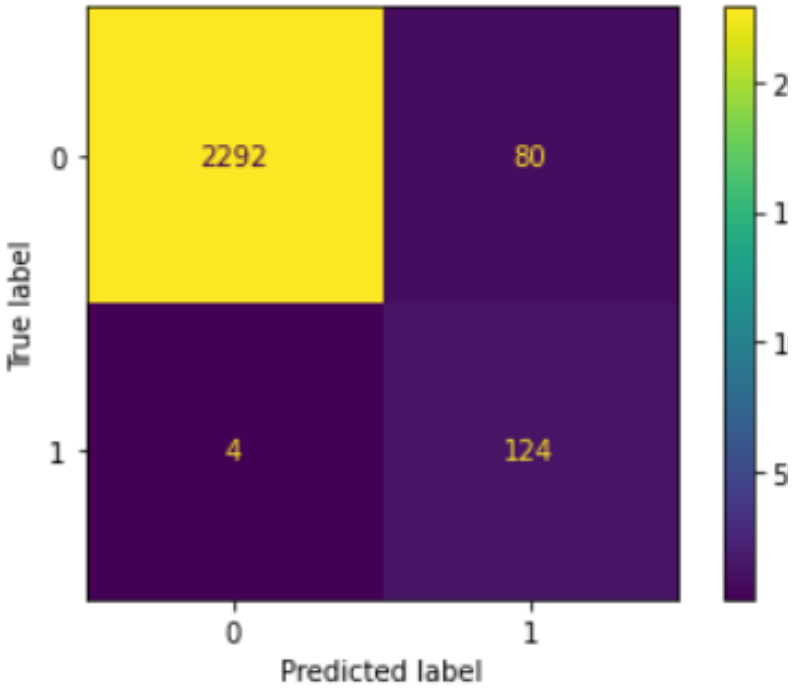
Resampled Models

- All three models had accuracy scores over 80%
- Adaboost had the highest test accuracy of 97%
- Multinomial Naïve bayes had lowest test accuracy of 82%
- Adaboost performed the best out the three resampled models

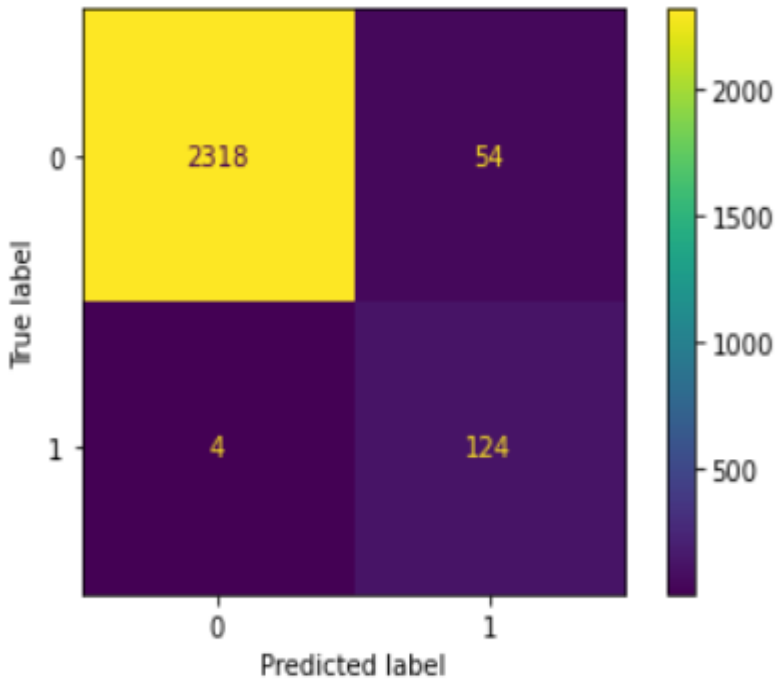
Multinomial Naïve Bayes



K-nearest neighbor



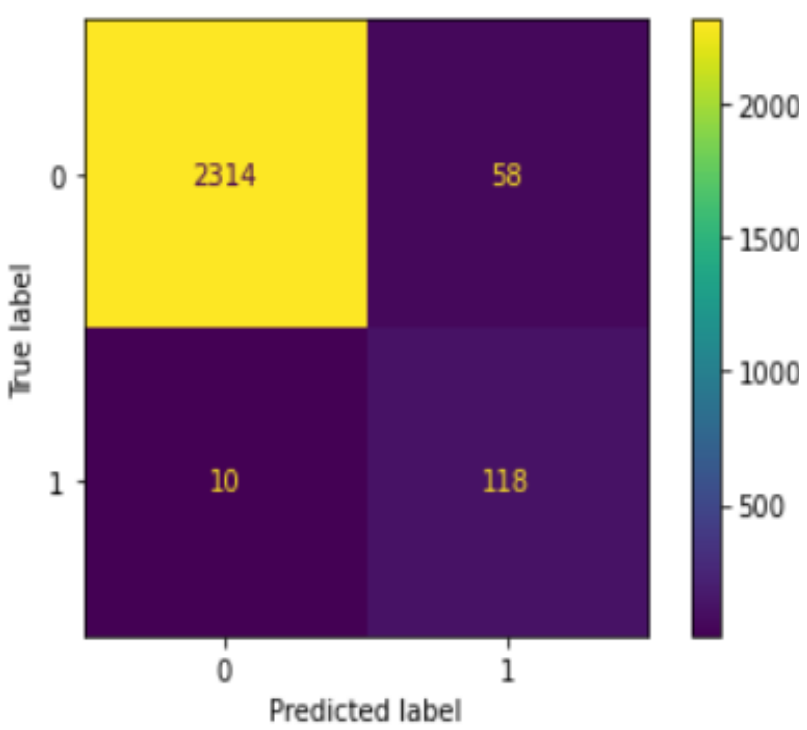
Adaboost w/Decision Tree



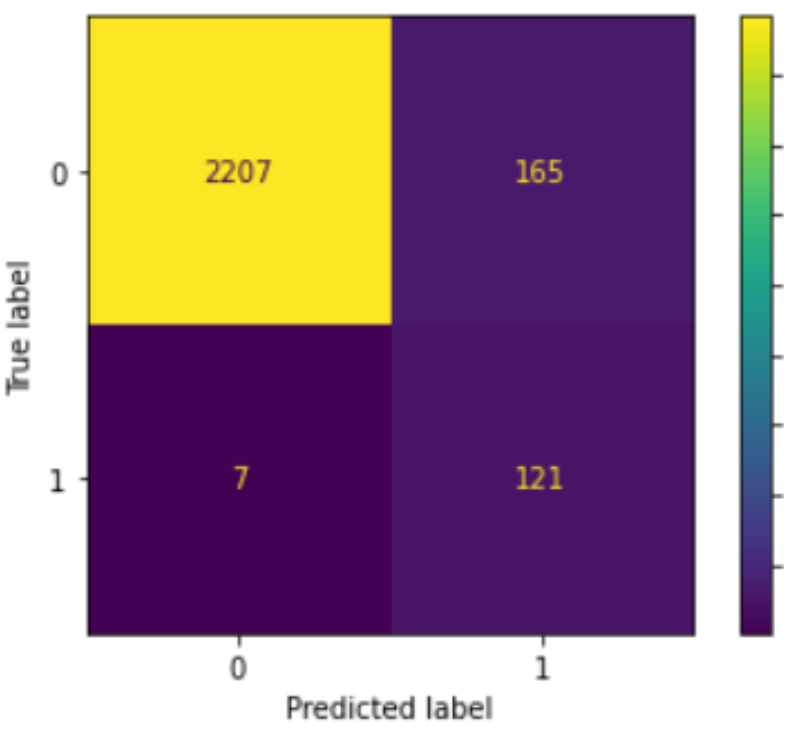
Regular Models

- All models fit directly on the training data had accuracy scores over 90%
- XG Boost performed the best of the three with a 98% accuracy score
- XG Boost had the lowest misclassification rate and best precision score
- Random Forest performed the worst of the three with an accuracy of 93%

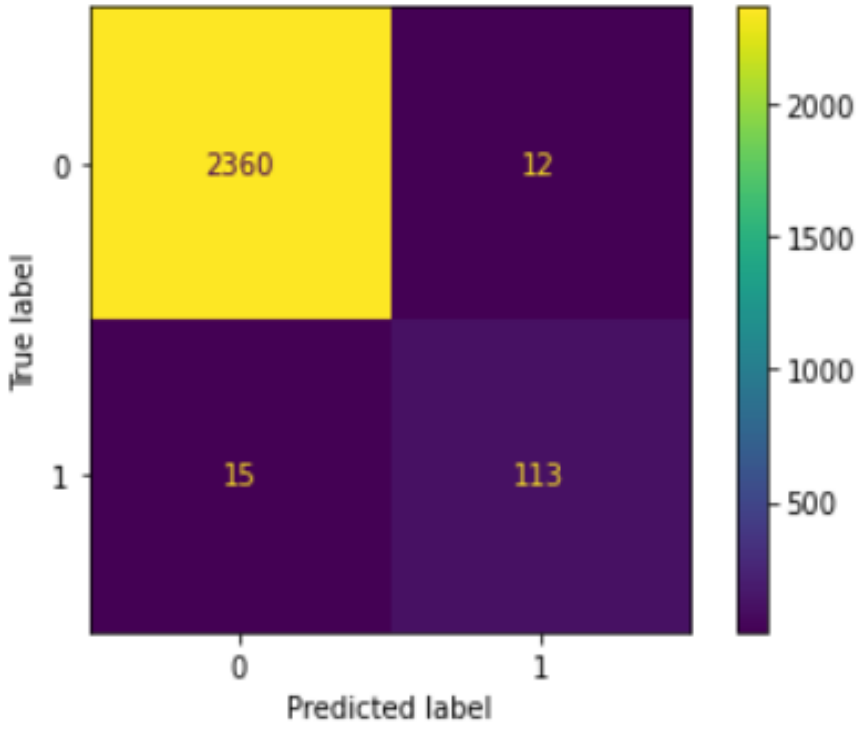
Logistic Regression



Random Forest



XG Boost



Conclusions and Recommendations

- All models had accuracy scores over 80% so a model was made that can predict fully paid loan status
- The best overall model was the XG Boost model
- Managerial roles may give the most likelihood of a fully paid loan status
- Moving forward more data could be gathered from branch Wells Fargo banks as most of the data came from banks in California, Texas, New York, and Florida
- Provided more data predicted whether the paid loan status vs the defaulting loan status could prove beneficial

	precision	recall	f1-score	support
0	0.99	0.99	0.99	2372
1	0.90	0.88	0.89	128
accuracy			0.99	2500
macro avg	0.95	0.94	0.94	2500
weighted avg	0.99	0.99	0.99	2500