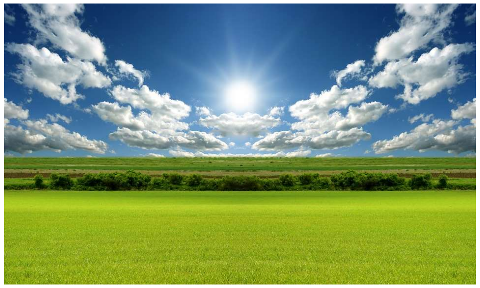
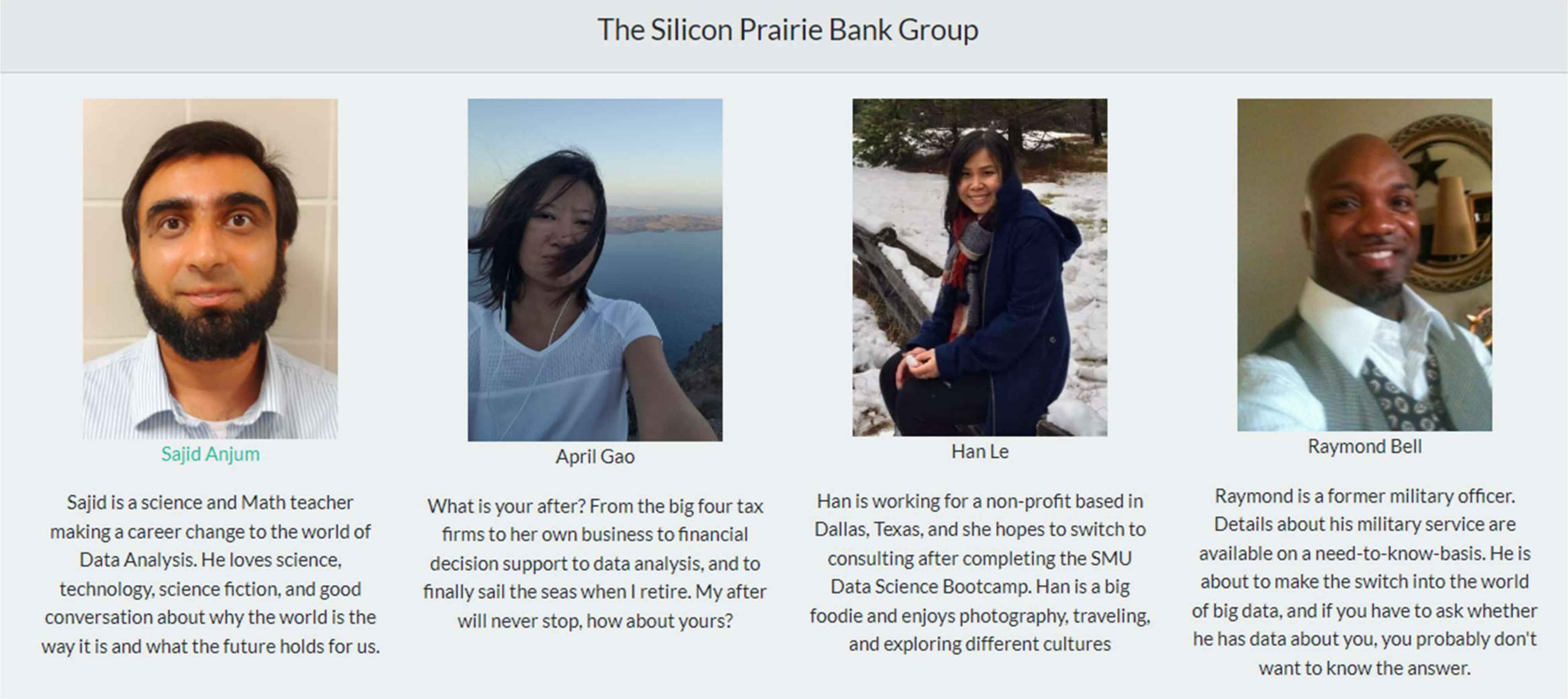
"To make a prairie it takes a clover and one bee,
One clover, and a bee.
And revery."

--Emily Dickinson

silicon prairie



The Bank Where Service Never Ends



The Silicon Prairie Bank Group

Sajid Anjum

Sajid is a science and Math teacher making a career change to the world of Data Analysis. He loves science, technology, science fiction, and good conversation about why the world is the way it is and what the future holds for us.

April Gao

What is your after? From the big four tax firms to her own business to financial decision support to data analysis, and to finally sail the seas when I retire. My after will never stop, how about yours?

Han Le

Han is working for a non-profit based in Dallas, Texas, and she hopes to switch to consulting after completing the SMU Data Science Bootcamp. Han is a big foodie and enjoys photography, traveling, and exploring different cultures

Raymond Bell

Raymond is a former military officer. Details about his military service are available on a need-to-know-basis. He is about to make the switch into the world of big data, and if you have to ask whether he has data about you, you probably don't want to know the answer.

Introduction

Twentieth Century Banking

- Prior to the advent of modern banking, loans were only accessible to a select and connected elite.
- The development of the Federal Reserve, ATMs, credit cards, and internet banking has allowed banking and financial services to be accessible to the majority of the American population
- Stability without centralization is the goal of a dynamic and prosperous financial system, and this is only possible if banks can make good decisions that minimize risk.

How Can We Make Banking Better?

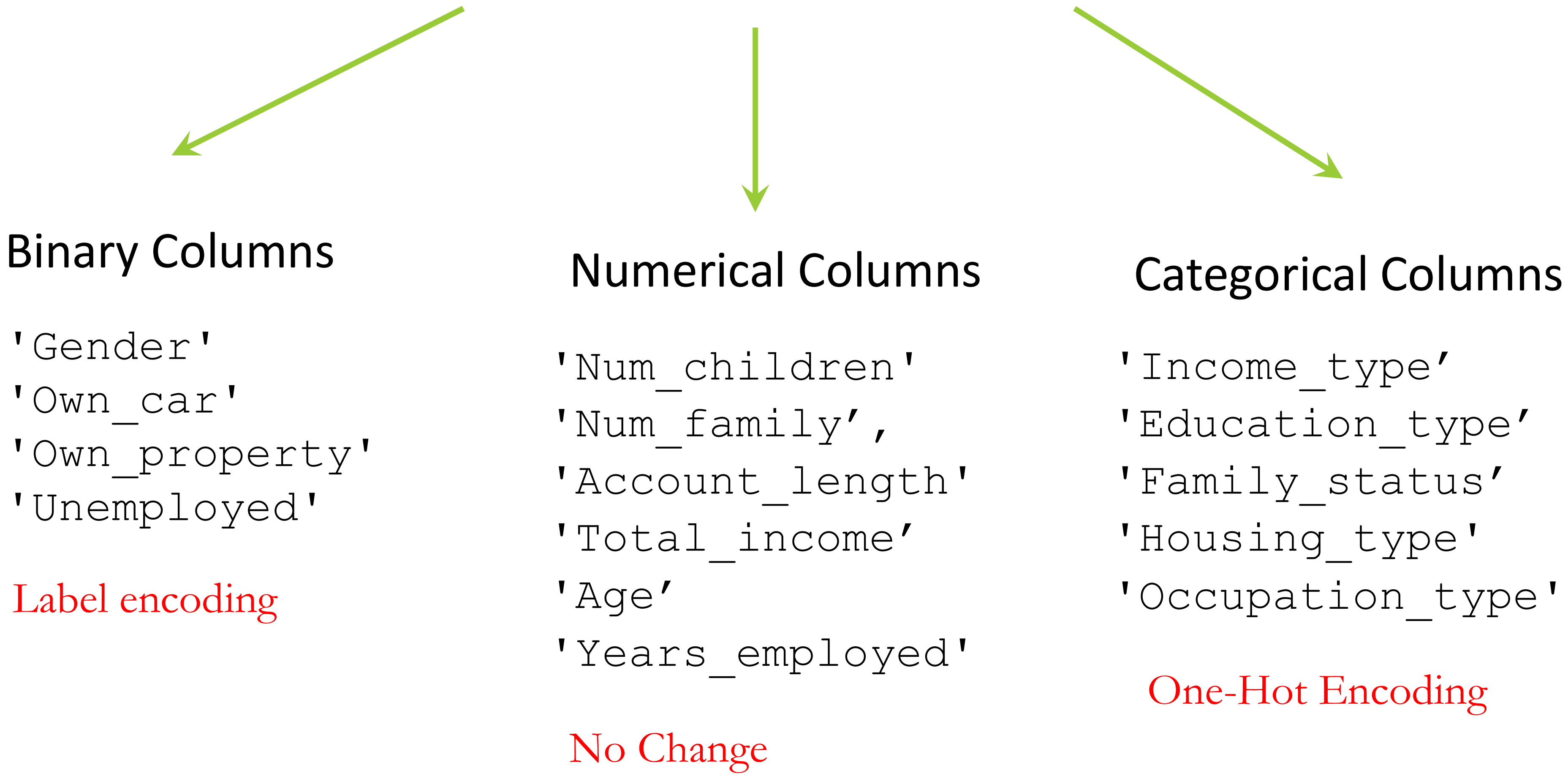
- We created two machine learning models and a web app to deploy them
- The first model helps us to predict whether an applicant will default on their credit card
- The second model helps banks to predict whether an applicant will default on their personal loan.
- We hope that helping banks make better decisions will allow them to customize credit cards and loans for every single applicant.

Data Processing and Machine Learning

Data Sets

- We obtained both of our datasets from Kaggle
- Our credit card data set was a dataset of customers in China who had applied for a credit card
- The dataset was very clean with no null values, binary columns already encoded as 0s and 1s, and numerical columns in appropriate datatypes.
- We dropped columns that referenced whether the applicant had a phone, work phone, or email address, as we did not think those columns would have predictive power.

Preparation of Data



Binary Columns

'Gender'
'Own_car'
'Own_property'
'Unemployed'

Label encoding

Numerical Columns

'Num_children'
'Num_family’,
'Account_length'
'Total_income’
'Age’
'Years_employed'

No Change

Categorical Columns

'Income_type’
'Education_type’
'Family_status’
'Housing_type'
'Occupation_type'

One-Hot Encoding

Preparation of Data

Target Column: **Approved or Denied** Binary encoded

$$\begin{aligned} \mathbf{Approved} &= \mathbf{0} \\ \mathbf{Denied} &= \mathbf{1} \end{aligned}$$

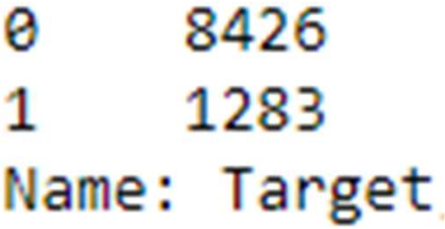
Our final dataset had fifty columns excluding the Target.

We felt that classification machine learning algorithms would be best for this sort of data set

Preparation of Data

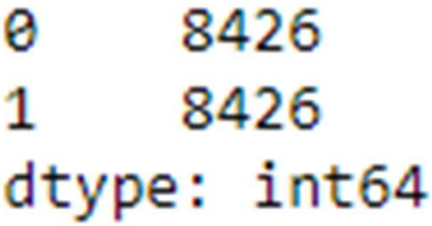
Use SMOTE to oversample the data with $\mathbf{Target} = \mathbf{1}$

Imbalance



```
0    8426
1    1283
Name: Target
```

Balance



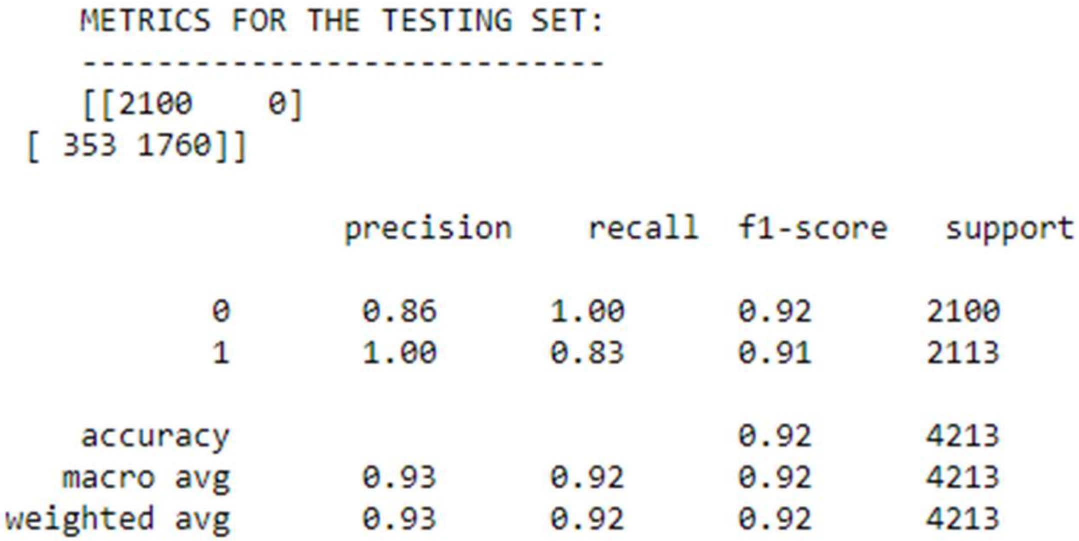
```
0    8426
1    8426
dtype: int64
```

Machine Learning

- A Tree model (LGBMClassifer) worked best

- The model was able to identify every single default

- The recall was 0.83, which means that for every 5 loans the model correctly identified as a default, one good customer was incorrectly predicted to default.

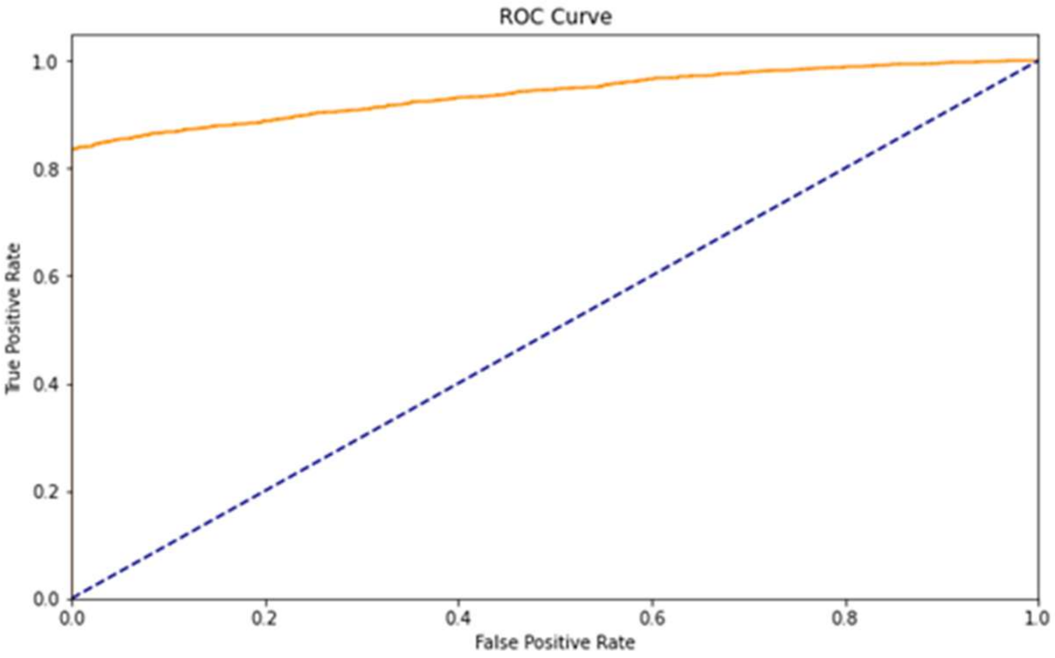


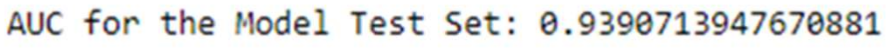
```
     METRICS FOR THE TESTING SET:
     ----------------------------
     [[2100    0]
 [ 353 1760]]

              precision    recall  f1-score   support

           0       0.86      1.00      0.92      2100
           1       1.00      0.83      0.91      2113

    accuracy                           0.92      4213
   macro avg       0.93      0.92      0.92      4213
weighted avg       0.93      0.92      0.92      4213
```

Machine Learning

- The AUC for the ROC curve was about 94%.

- We conclude that the model will be great at identifying defaults, although perhaps overpredicting defaulters just a tad.





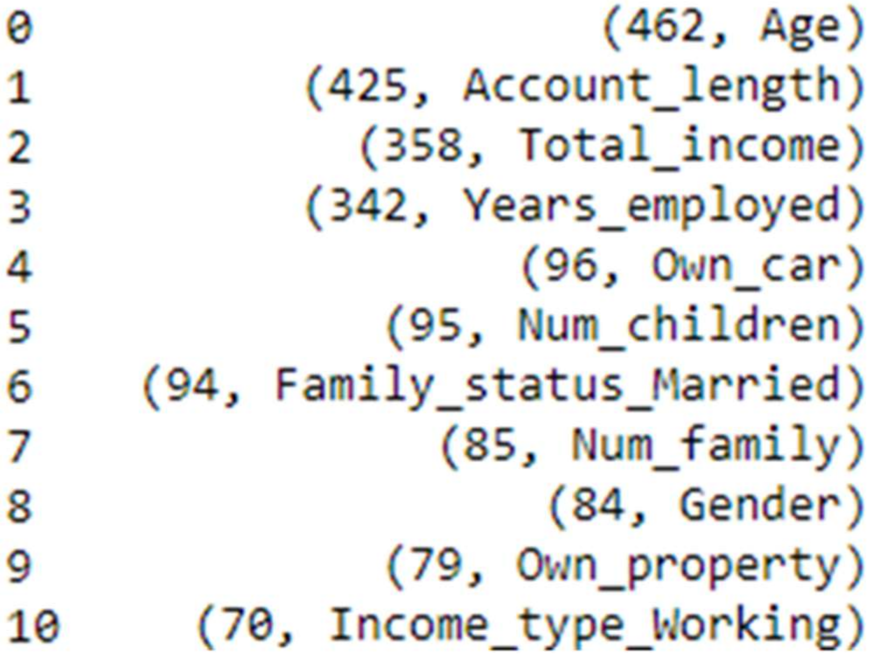
```
AUC for the Model Test Set: 0.9390713947670881
```

Machine Learning (Explainability)

- The most important features of our model were:

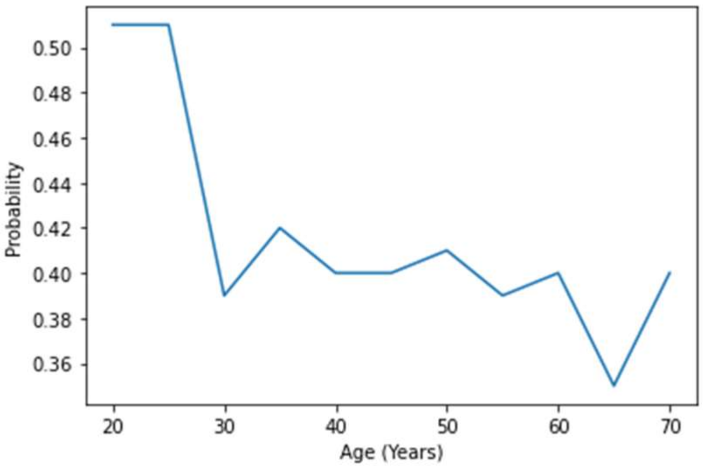
- We decided to see how the probability of our model predictions varied versus a variation in these features.

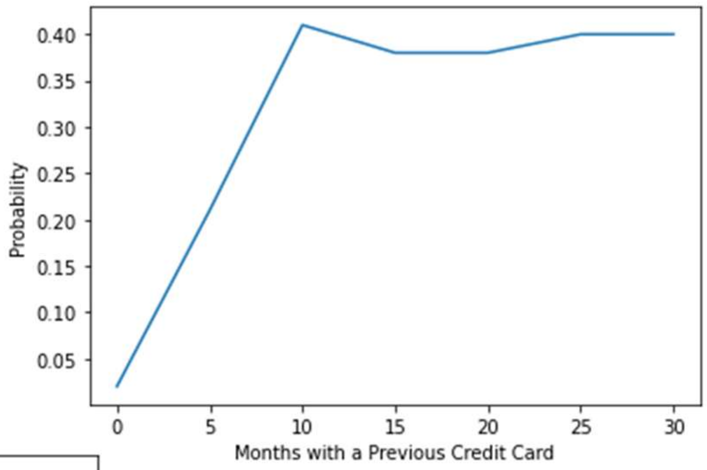
- We created a fake “average” customer and varied these features.

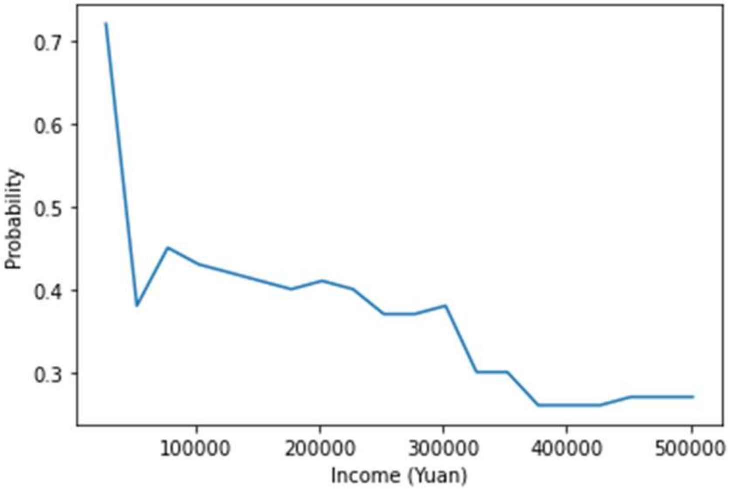


```
0                      (462, Age)
1           (425, Account_length)
2             (358, Total_income)
3           (342, Years_employed)
4                  (96, Own_car)
5             (95, Num_children)
6    (94, Family_status_Married)
7               (85, Num_family)
8                   (84, Gender)
9             (79, Own_property)
10     (70, Income_type_Working)
```

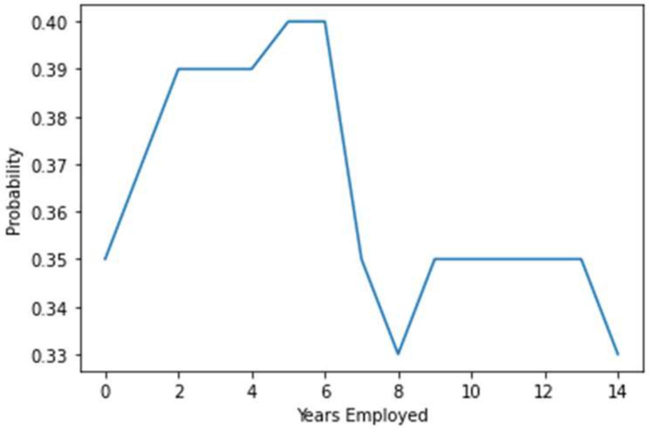
Machine Learning (Explainability)

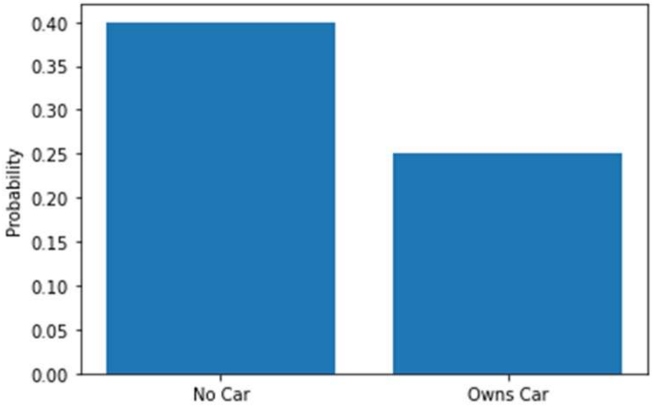


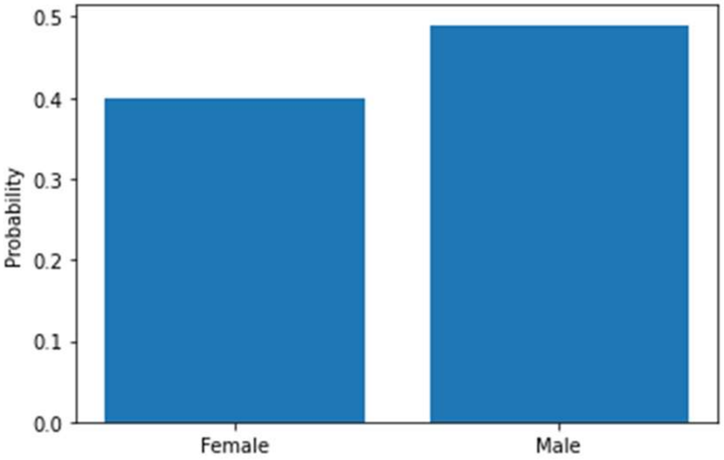




Machine Learning (Explainability)



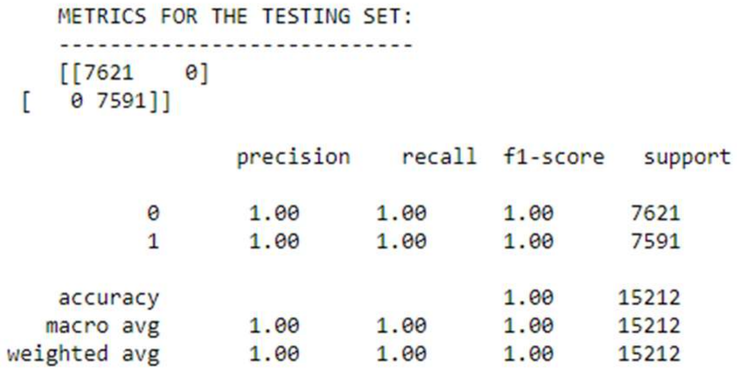




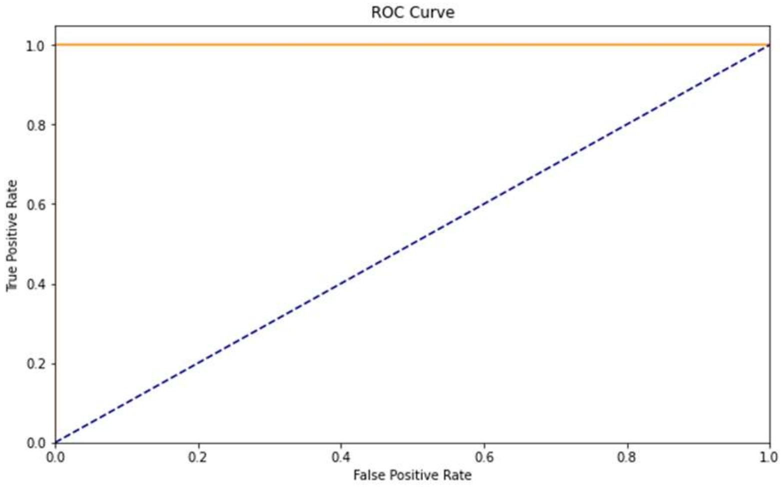
Machine Learning

- We used a RandomForestClassifier to create a model for our loan dataset. The details are in our writeup. Unfortunately, we were not able to deploy the model on our website.

- This is a shame as the model was perfect!



```
    METRICS FOR THE TESTING SET:
    ----------------------------
    [[7621    0]
 [    0 7591]]

                  precision    recall  f1-score   support

             0       1.00      1.00      1.00      7621
             1       1.00      1.00      1.00      7591

      accuracy                           1.00     15212
     macro avg       1.00      1.00      1.00     15212
  weighted avg       1.00      1.00      1.00     15212
```



Website

Conclusions

Conclusions

- The tree models worked the best. For credit cards, both the LGBMClassifier and the RandomForestClassifier gave spectacular results, but we picked the LGBMClassifier because it showed less evidence of overfitting.

- Age, number of months with a previous credit card, total income, and total years of employment are, not surprisingly, the categories that had the largest impact on the approval algorithm

- Our model made predictions similar to the patterns that we saw in the Tableau Dashboards

Limitations and Future Work

Limitations

- The biggest limitation was time. We were not able to implement our second model, and there were so many interesting patterns that we did feel we did not identify.

- If we could find location data for credit card, we feel that it would provide a lot of explanatory power. It would be interesting how different regions of the nation approach financing.

- The machine learning model trained very quickly on our data. We feel we could process more data and perhaps make a better model.

Acknowledgments

We would like to thank Professor Alexander Booth, Professor Farzad HosseinAli and SMU for giving us a wonderful bootcamp experience over the past six months. We would also thank all the other members of our cohort for creating a friendly and collaborative learning environment.