Credit Card Default

Technical Presentation

By Eric Wheeler

Problem Introduction

In order to mitigate risk in the financial industry, it is common to use personal information and data submitted by credit card applicants to objectively quantify risk and predict the probability of default. This provides banks with the ability to decide whether to issue a credit card to an applicant in the future.

DELINQUENCIES

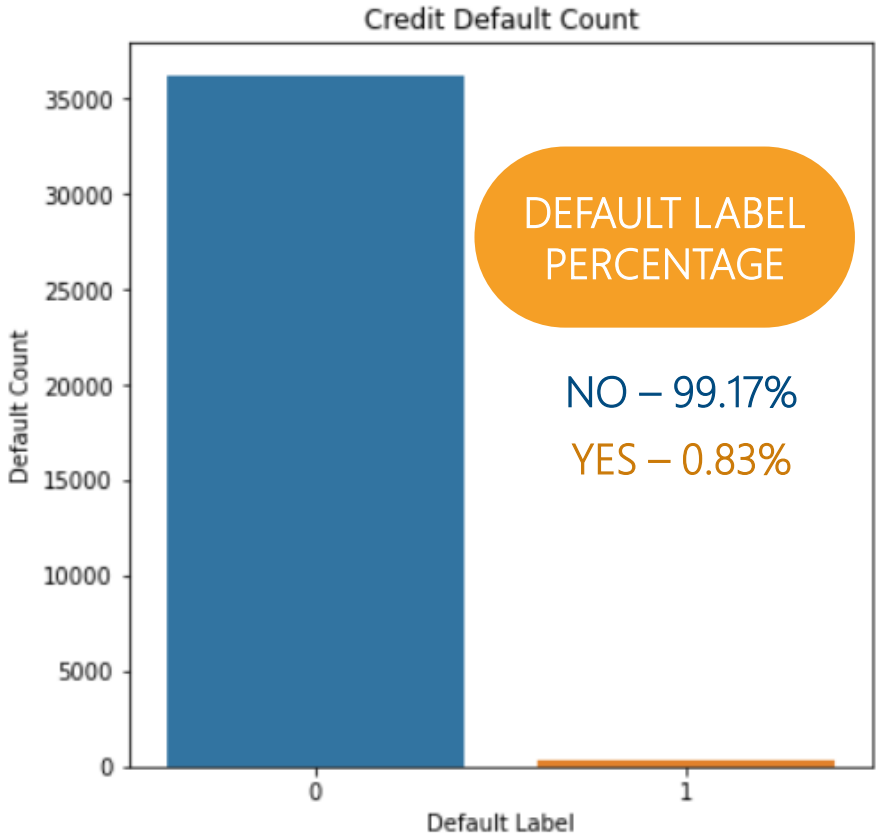
The American Bankers Association defines a delinquency as a late payment that is 30 days or more overdue. [1]

DEFAULT LABEL

TransUnion's Industry Insights Report found the credit card delinquency rate reached 1.22% in Q3 2020. This figure is based on accounts that are 90 days or more overdue. [2]

DEFAULT

After an individual has failed to make a payment on their credit card for 180 days, the account is considered to be in default. [3]



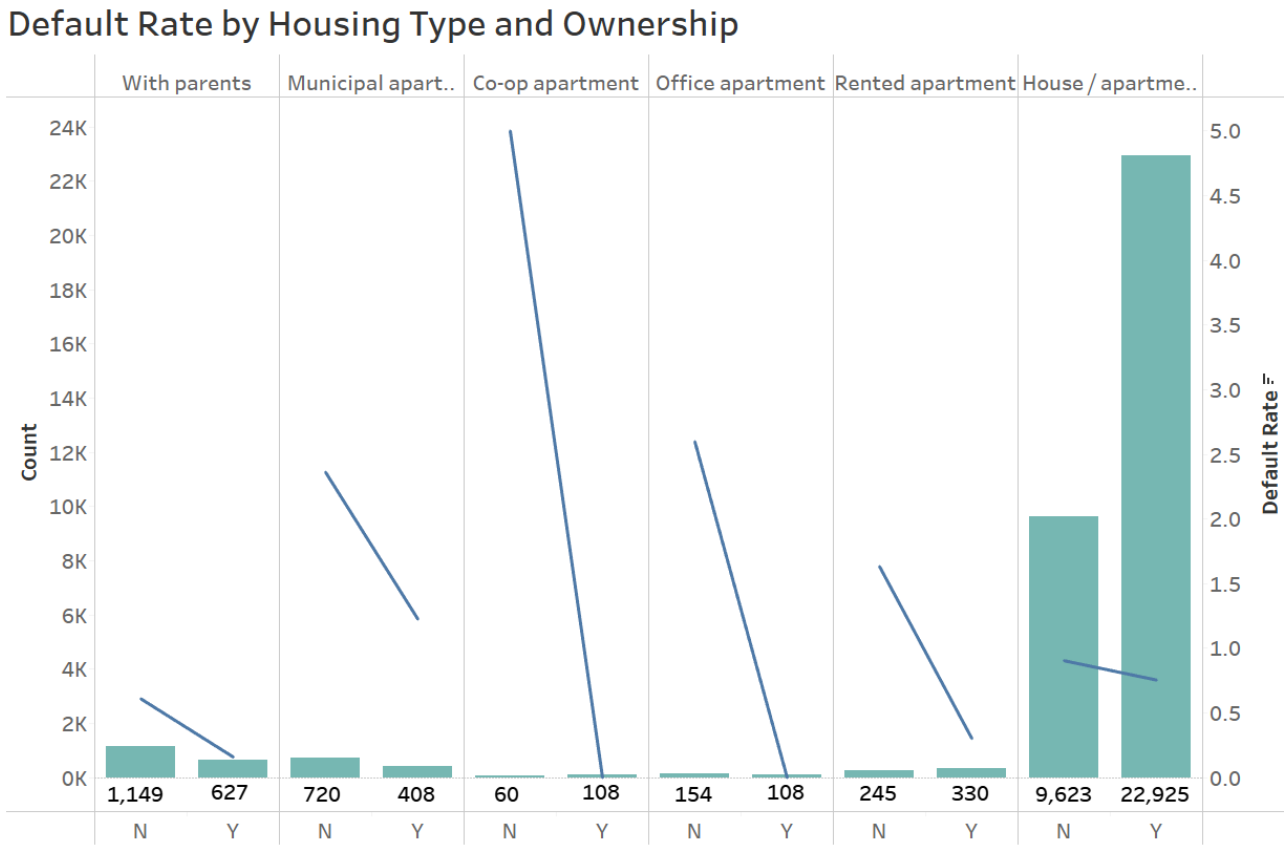
1. 1. ABA Report: Closed-End Loan Delinquencies Rise in First Quarter at Onset of COVID-19
2. Credit Card Balances at Lowest Levels Since 2017; Holiday Season Credit Usage in the Spotlight
3. What You Can Do About Credit Card Default?

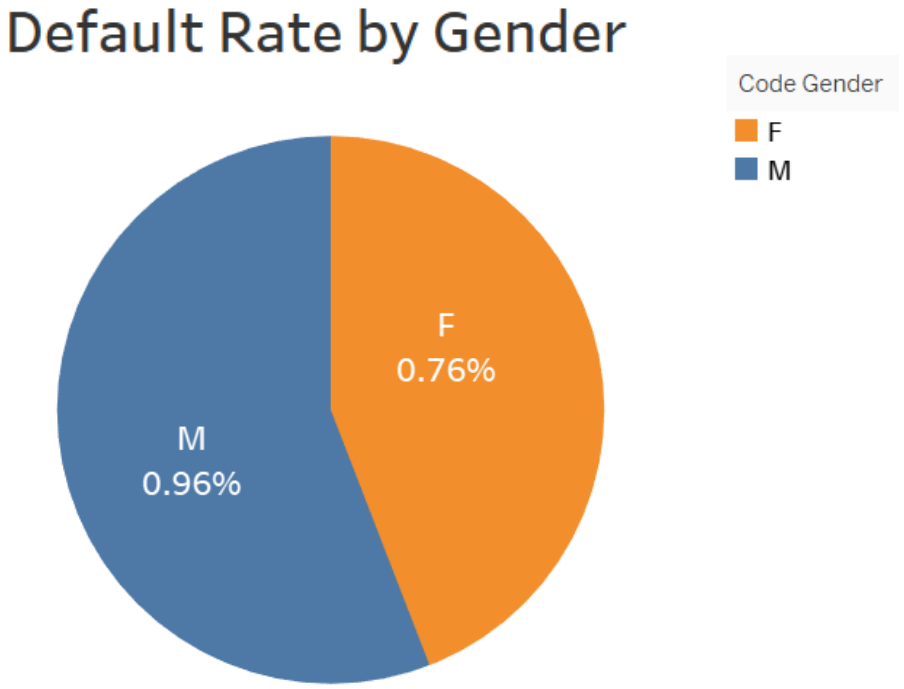
Technical Approach

Can we use banking information, past credit history, and statistical techniques to mitigate the risk to financial institutions by determining the factors that contribute to credit default?

ISSUES TO EXPLORE	PURPOSE
Exploratory Data Analysis	To determine the significant variables that contribute towards default
Logistic Regression Model	Suitable for binary classification tasks

Exploratory Data Analysis





Property Ownership

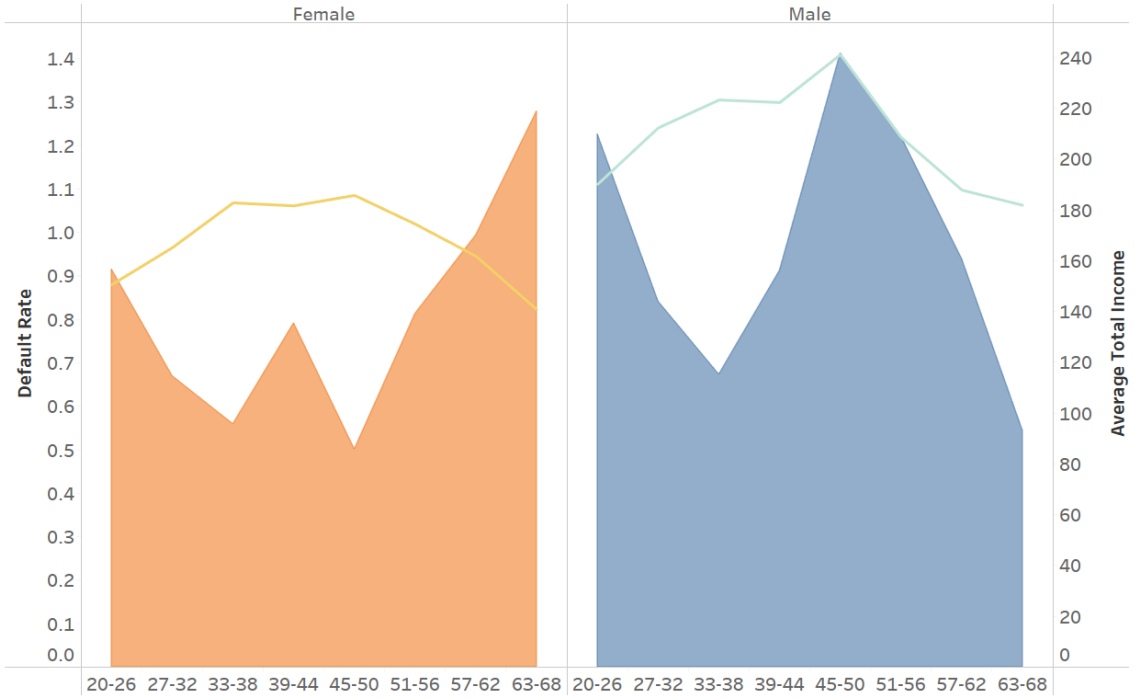
Individuals who do not own their property are more likely to default than those who do own, regardless of their housing type. This was especially prominent amongst lower income housing and shared housing.

Gender

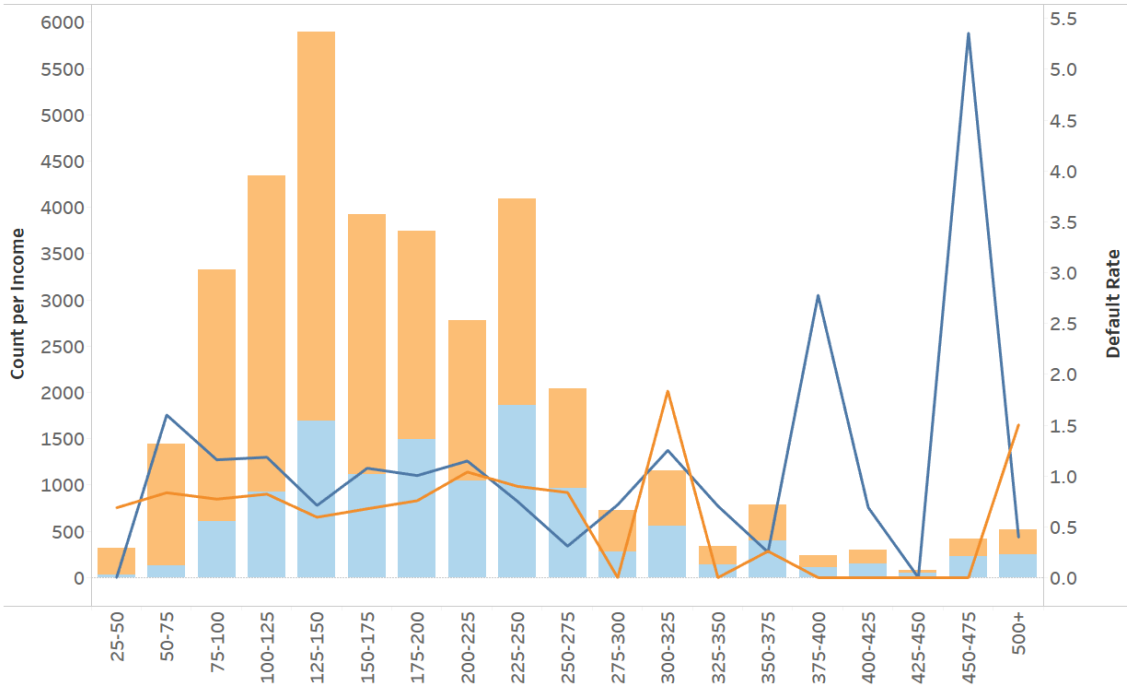
The data shows that men have a slightly higher likelihood of default when compared with women.

Exploratory Data Analysis

Default Rate and Average Income by Age



Default Rate by Income



Average Income by Age

Looking across age groups, it appears that for both men and women, average income increases up until the ages 45-50 and then begins to drop. Interestingly, during this period of decreasing wages, women appear to increase their likelihood of default whereas men appear to begin defaulting at a lower rate.

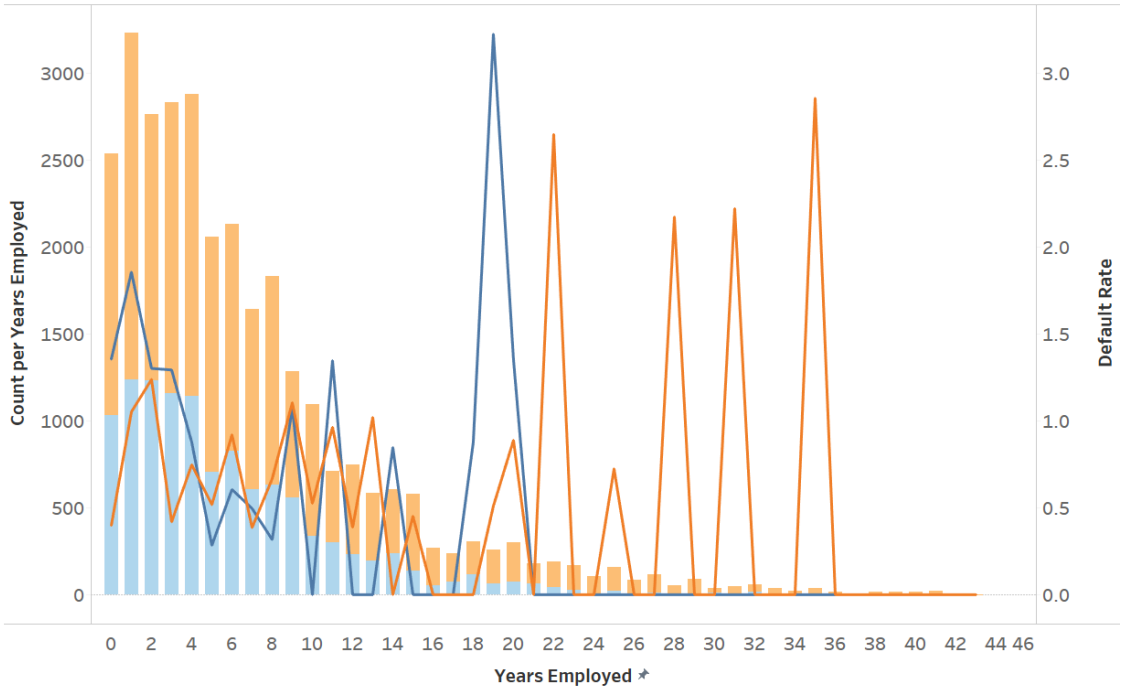
Code Gender
F
M

Total Income

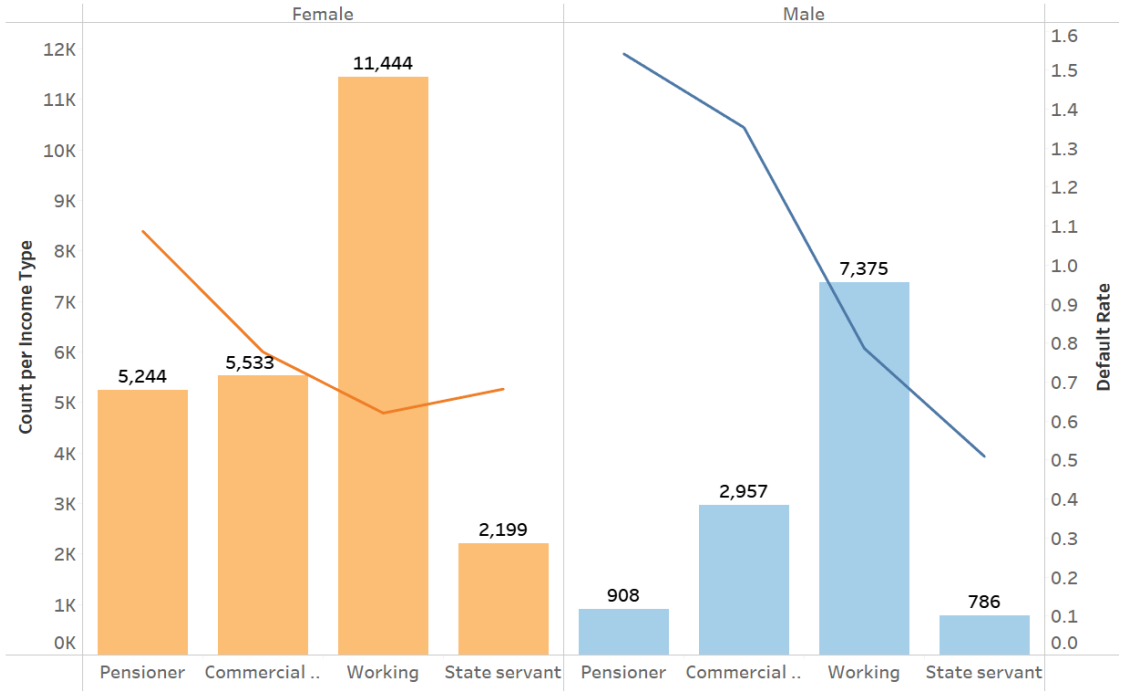
While women seem to have the highest default rate when they are oldest, men appear to default the most when they are at the peak of their earning potential and have an income larger than $375k.

Exploratory Data Analysis

Default Rate by Years Employed



Default Rate by Income Type



Code Gender
F
M

Years Employed

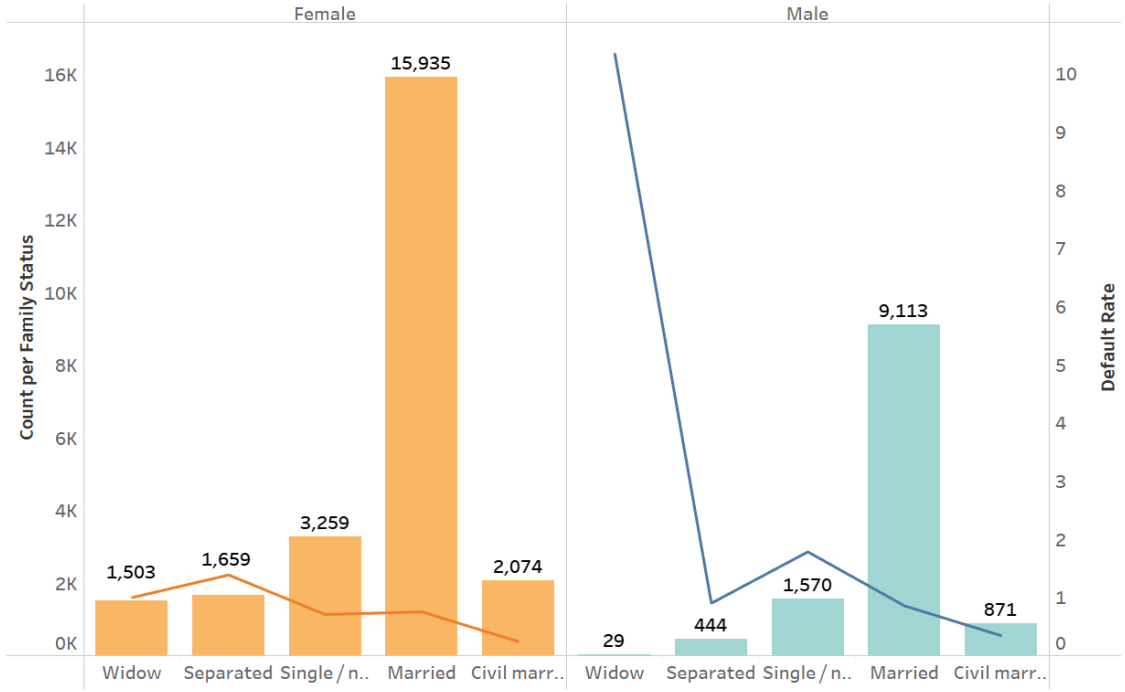
For women in particular, longer tenure in their careers seems to coincide with a higher default rate, presumably due to a decrease in wages brought on by increased age.

Income Type

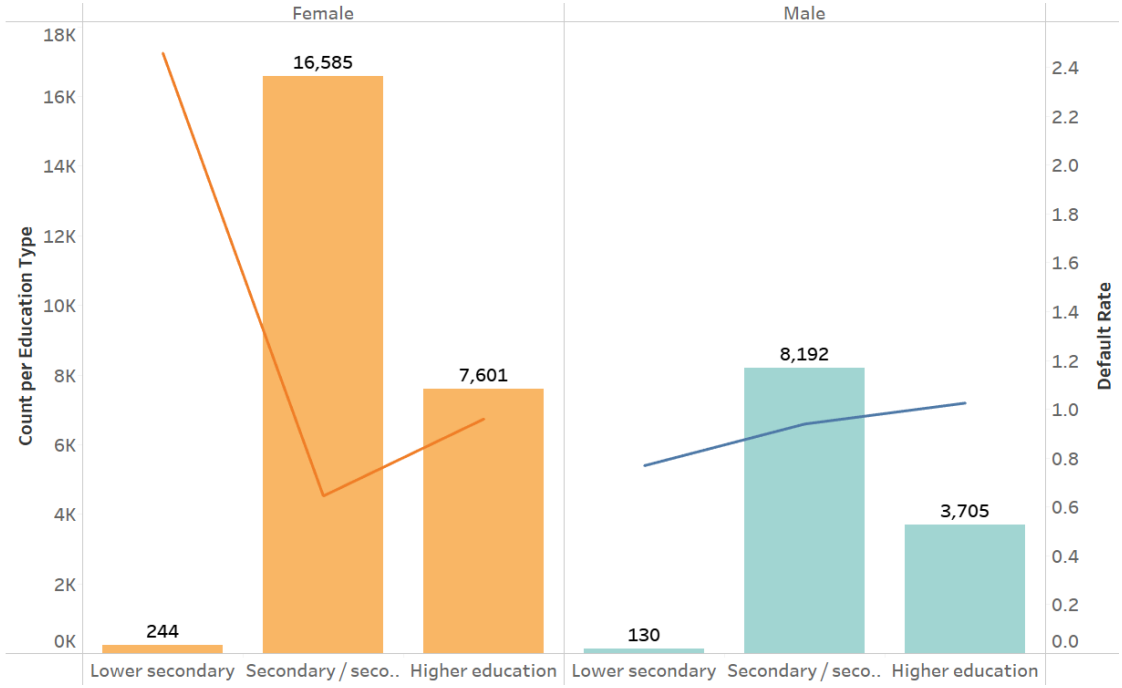
This trend seems to continue until retirement seeing as both male and female pensioners have a much higher default rate when compared to working individuals.

Exploratory Data Analysis

Default Rate by Family Status



Default Rate by Education Level



Code Gender
F
M

Family Status

Individuals with a family status of single, including those as a result of separation or loss, have a higher default rate than married individuals.

Education Level

Surprisingly, men have the highest default rate when they have a higher education whereas for women it is when they have the lowest education.

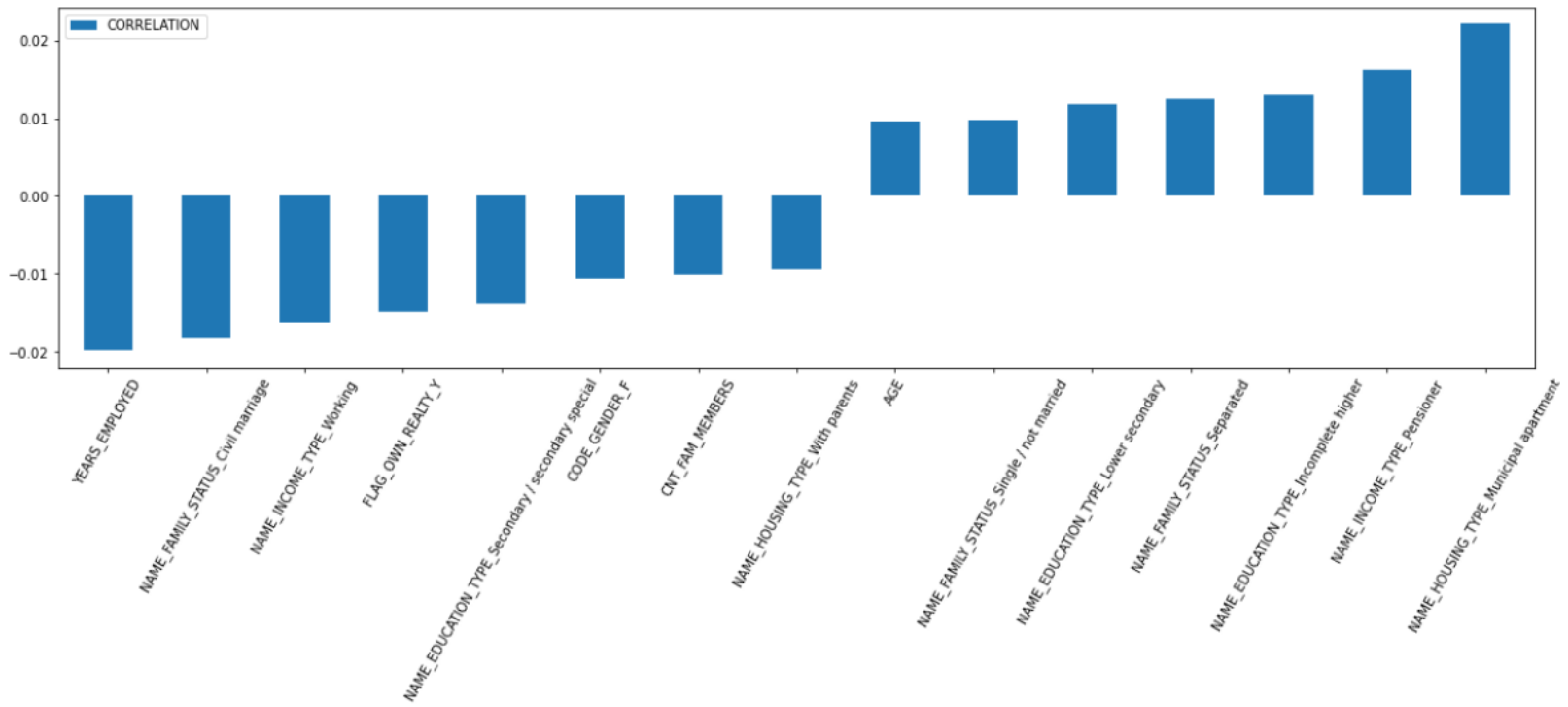
Feature Selection

	INDEP_VAR	CORRELATION
0	YEARS_EMPLOYED	-0.019833
1	NAME_FAMILY_STATUS_Civil marriage	-0.018209
2	NAME_INCOME_TYPE_Working	-0.016285
3	FLAG_OWN_REALTY_Y	-0.014829
4	NAME_EDUCATION_TYPE_Secondary / secondary special	-0.013779
5	CODE_GENDER_F	-0.010538
6	CNT_FAM_MEMBERS	-0.009936
7	NAME_HOUSING_TYPE_With parents	-0.009436
8	NAME_HOUSING_TYPE_House / apartment	-0.009409
9	NAME_INCOME_TYPE_State servant	-0.006321
10	NAME_FAMILY_STATUS_Married	-0.006194
11	FLAG_OWN_CAR_Y	-0.004784
12	CNT_CHILDREN	-0.004708
13	NAME_EDUCATION_TYPE_Academic degree	-0.002709
14	NAME_INCOME_TYPE_Student	-0.001588
15	NAME_HOUSING_TYPE_Rented apartment	0.000575
16	AMT_INCOME_TOTAL	0.003103
17	FLAG_OWN_CAR_N	0.004784
18	NAME_EDUCATION_TYPE_Higher education	0.006328
19	NAME_HOUSING_TYPE_Office apartment	0.006555
20	NAME_HOUSING_TYPE_Co-op apartment	0.007187
21	NAME_FAMILY_STATUS_Widow	0.008008
22	NAME_INCOME_TYPE_Commercial associate	0.009073
23	AGE	0.009501
24	NAME_FAMILY_STATUS_Single / not married	0.009818
25	CODE_GENDER_M	0.010538
26	NAME_EDUCATION_TYPE_Lower secondary	0.011719
27	NAME_FAMILY_STATUS_Separated	0.012434
28	NAME_EDUCATION_TYPE_Incomplete higher	0.013058
29	FLAG_OWN_REALTY_N	0.014829
30	NAME_INCOME_TYPE_Pensioner	0.016192
31	NAME_HOUSING_TYPE_Municipal apartment	0.022119

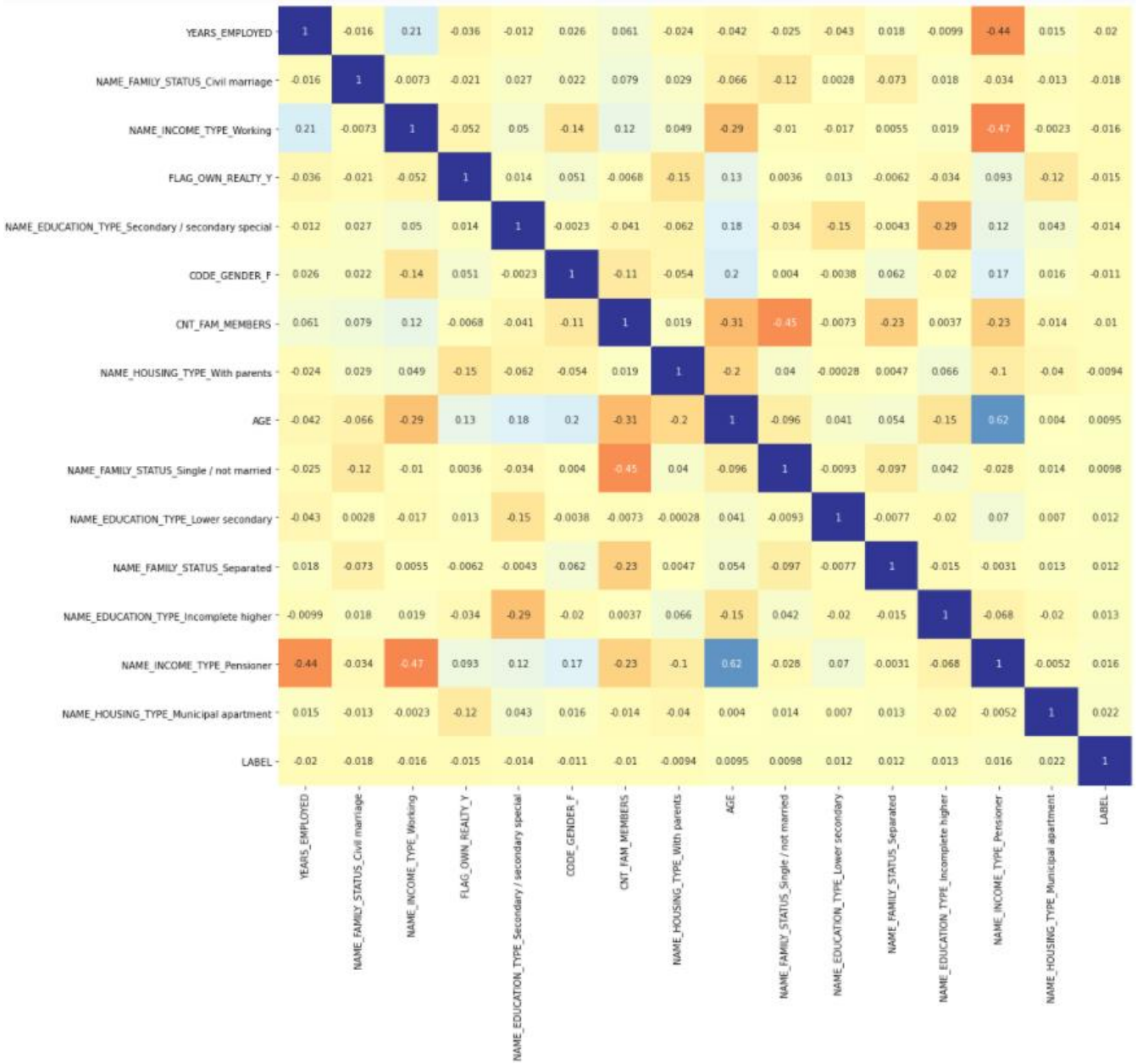
In order to utilize the categorical features for the model, they first needed to be transformed using dummy variables.

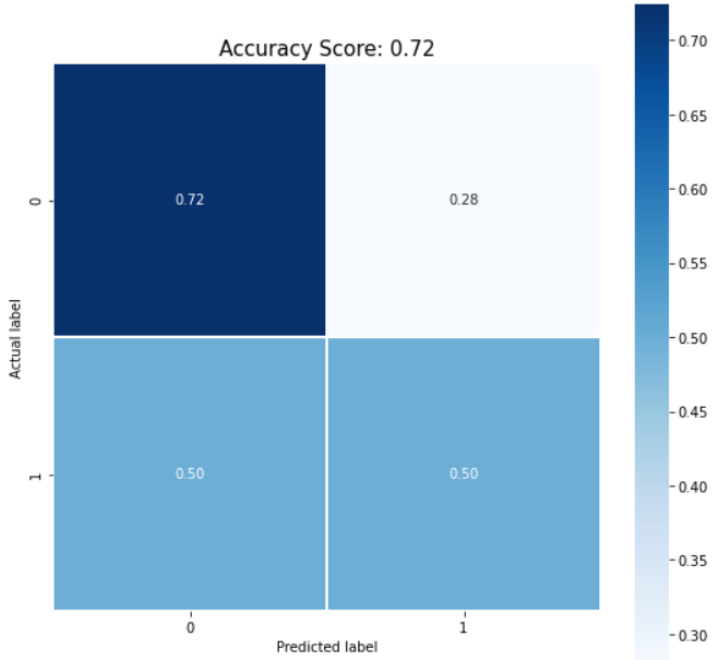
I then calculated the correlation between all the independent variables and the default label. This provided me with their correlation coefficients.

I chose the 15 most highly correlated features to build the model and tested its performance.



The Initial Model



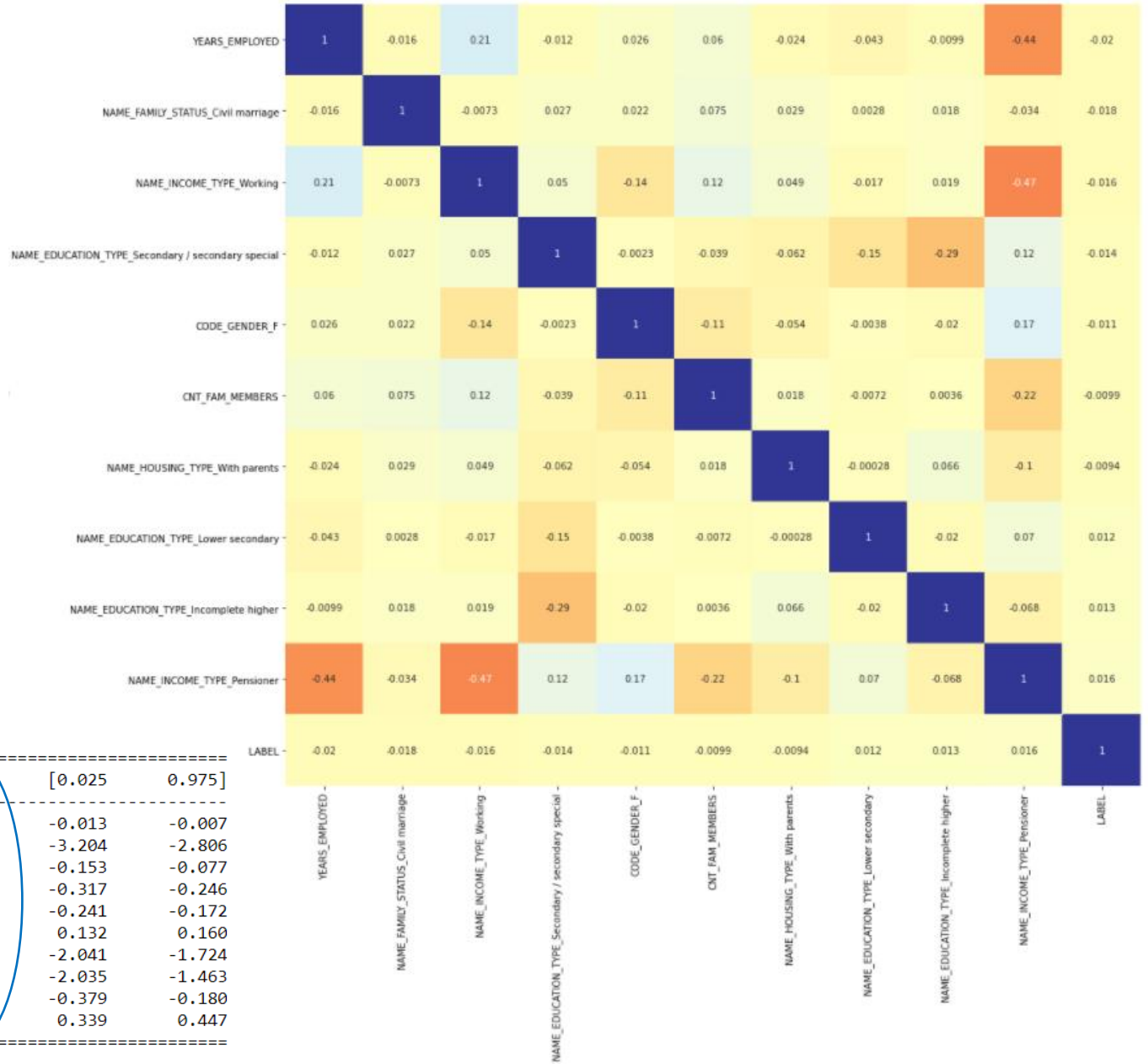


This resulted in a model with an accuracy score of 0.72, but a recall of only 0.5. In order to improve the model so that it might capture more than 50% of the customers at risk of default, a reduction of input features was necessary. I noticed that the age and pensioner variables had the strongest multicollinearity amongst the model and decided that one of these would be a good place to start.

Multicollinearity and Significance

To my surprise, the model responded more positively to the removal of the age variable than it did for the pensioner variable. After several rounds of testing and adjusting the model, I settled on these 10 variables as they proved to provide the most optimal results.

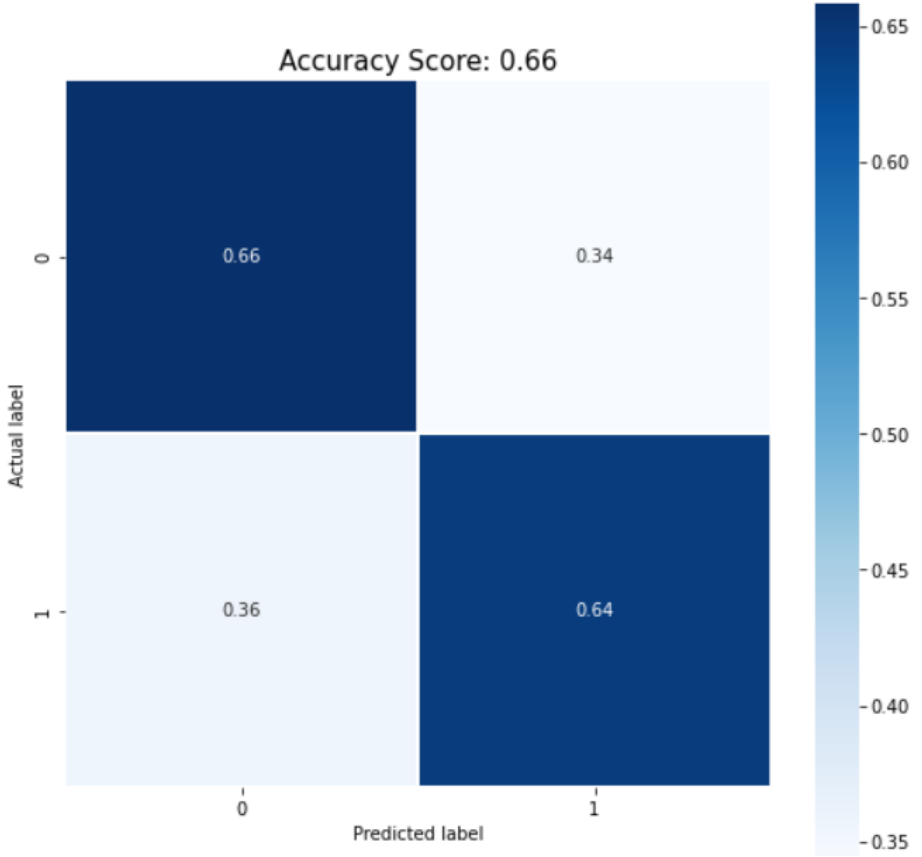
The feature selection was now complete as I found that their multicollinearity was acceptably low, and their p-values showed that they are all statistically significant.

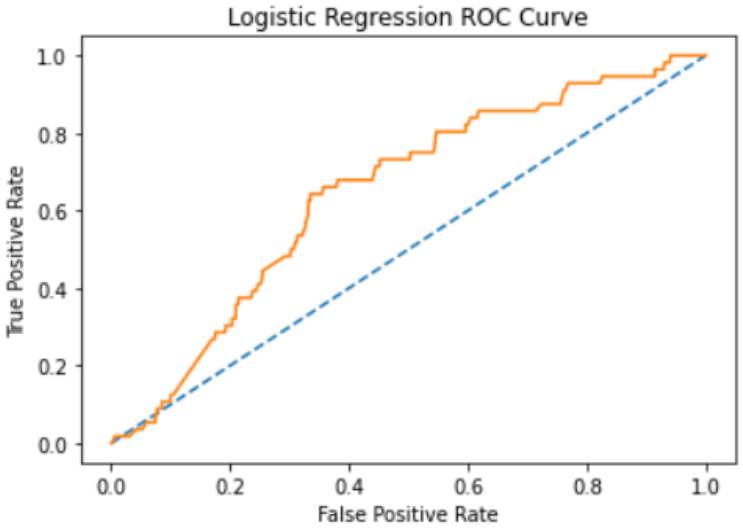


	coef	std err	z	P>|z|	[0.025	0.975]
YEARS_EMPLOYED	-0.0097	0.002	-6.305	0.000	-0.013	-0.007
NAME_FAMILY_STATUS_Civil marriage	-3.0053	0.101	-29.615	0.000	-3.204	-2.806
NAME_INCOME_TYPE_Working	-0.1150	0.019	-5.992	0.000	-0.153	-0.077
NAME_EDUCATION_TYPE_Secondary / secondary special	-0.2815	0.018	-15.506	0.000	-0.317	-0.246
CODE_GENDER_F	-0.2067	0.018	-11.793	0.000	-0.241	-0.172
CNT_FAM_MEMBERS	0.1458	0.007	19.977	0.000	0.132	0.160
NAME_HOUSING_TYPE_With parents	-1.8824	0.081	-23.267	0.000	-2.041	-1.724
NAME_EDUCATION_TYPE_Lower secondary	-1.7493	0.146	-11.989	0.000	-2.035	-1.463
NAME_EDUCATION_TYPE_Incomplete higher	-0.2795	0.051	-5.508	0.000	-0.379	-0.180
NAME_INCOME_TYPE_Pensioner	0.3929	0.028	14.223	0.000	0.339	0.447

The Final Model

After splitting the data into train and test groups, I oversampled the training data using SMOTE[4] to account for the imbalance in the label values.





Ultimately, I chose to emphasize increasing the recall of the default label as opposed to the accuracy of the model. For this particular problem I was willing to sacrifice over-estimation of the default label resulting in greater false positives in order to capture the largest number of true positives possible. As a result, the model was able to predict 64% of the default labels in the test data.

4. Synthetic Minority Oversampling Technique is used to synthesize new examples from the minority class

Thank You

Eric Wheeler

917-562-3434

Ericwheeler.90@gmail.com