**Credit One**

**Credit Default Prediction Study**

**Melissa Perez**

**What concrete information can we derive from the data we have?**

We analyzed the 30,000 observations (customer data) provided to us by Credit One. If these observations are representative of Credit One’s customers, then the information gleaned from this data can provide Credit One with valuable insight into their customer base.

**Credit One Customer Demographics**

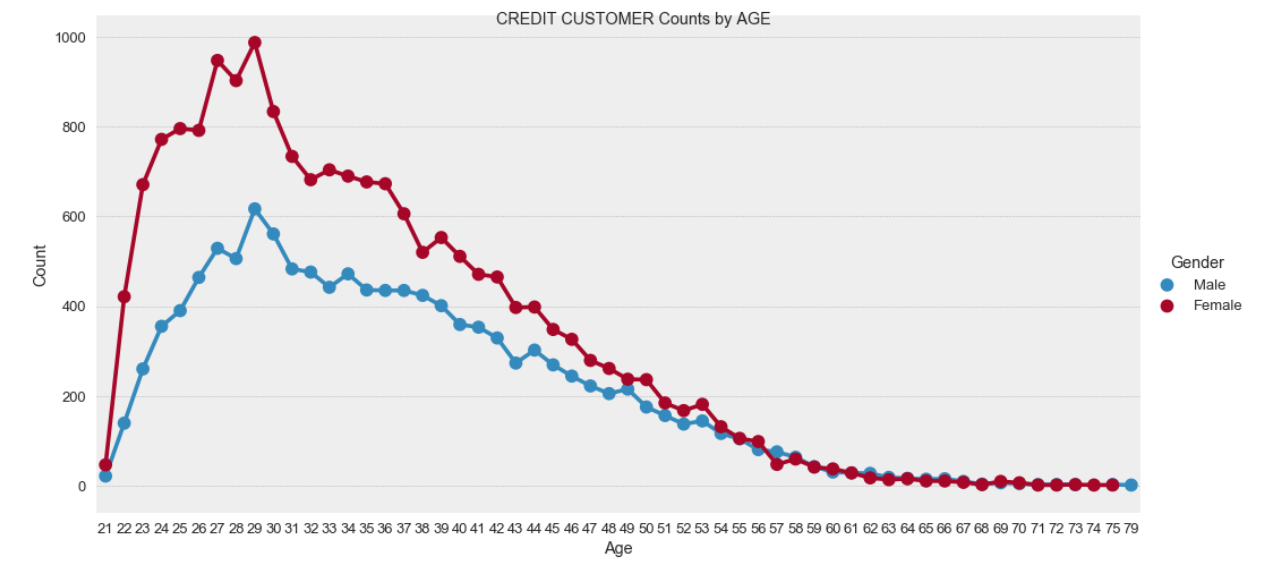
Credit One customers range in age from 21 to 79, with 78% of customers between 23 and 43.

The majority of Credit One customers are female.

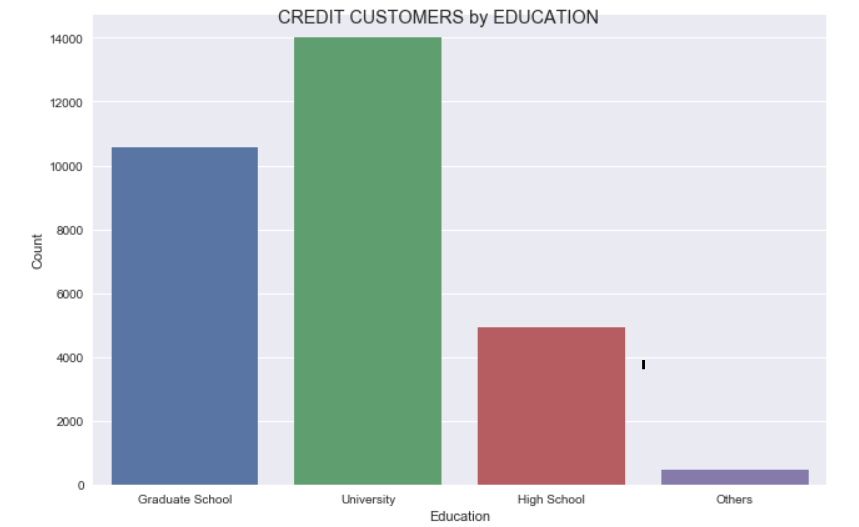
Gender: Female 60.37%

Male 39.63%

The following chart shows Credit One customers broken down by gender and plotted by age.

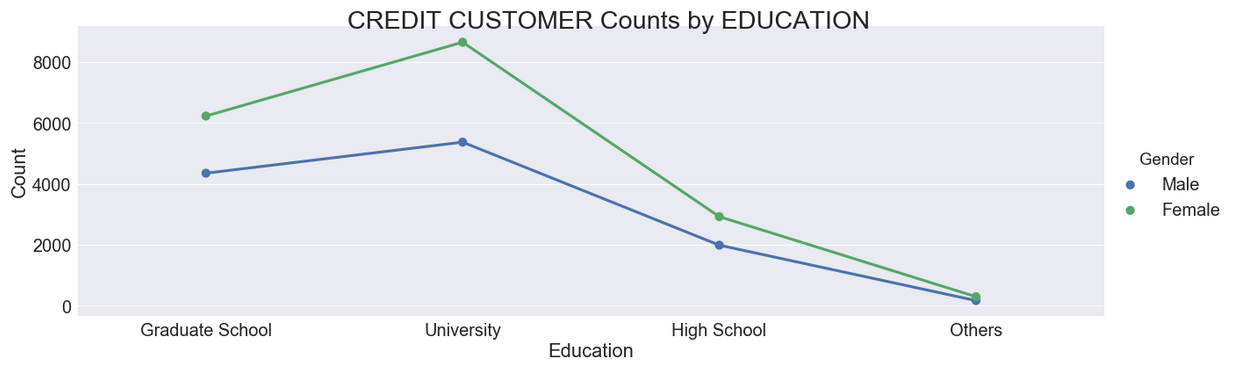


**Breakdown of Customers by Educational Level**

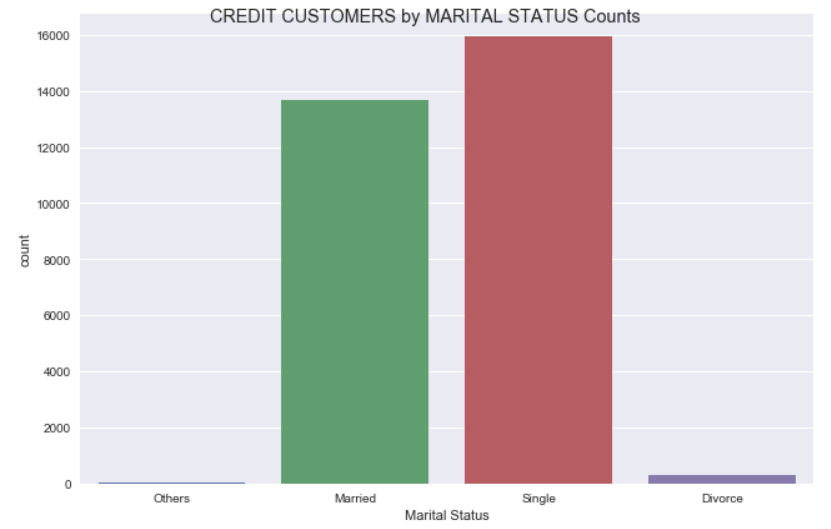


|  |  |
| --- | --- |
|  | **Percentage** |
| Graduate School | 35.28 |
| University | 46.77 |
| High School | 16.39 |
| Others | 1.56 |

The following chart shows Credit One customers broken down by gender and plotted by educational level.



**Breakdown of Customers by Marital Status**



The highest percentage customers are Single and Married.

Very few customers fall into the Divorced or Other categories.

**Breakdown of Customers who Defaulted by Gender**

|  |  |  |
| --- | --- | --- |
|  | **No-Default** | **Default** |
| Male | 9015 | 2873 (24.1%) |
| Female | 14349 | 3763 (20.8%) |

Interestingly, the data shows that, though fewer men than women acquired a loan, a higher percentage of those men defaulted relative to the women borrowers.

**How do you ensure that customers can/will pay their loans?**

We can’t ensure or guarantee that customers will repay their loans; however, we can attempt to predict which customers are more likely to repay their loans and which ones are not. That is the objective of this effort, to predict, using the data provided to us by Credit One, which customers are more likely to repay their loans.

**Can we approve customers with high certainty?**

After analyzing, cleaning, and prepping the data in *default of credit card files.csv,* we developed five models to predict customer default based on that data. Of the five classifiers, Random Forest predicted with the highest accuracy of 82.33%.

|  |  |
| --- | --- |
| **Classifier** | **Prediction Accuracy** |
| Random Forest | 82.33% |
| Decision Tree | 81.00% |
| K-Nearest Neighbor | 80.66% |
| .Logistic Regression | 79.66% |
| Gaussian NB | 38.66% |

When we look at the accuracy percentages of our prediction models, we consider them in relation to a baseline accuracy that can be achieved by always predicting the most frequent result. This is known as “null accuracy”.

For example, our loan default data shows that 77.8% of Credit One’s customers did not default on their loan. Seventy-seven point eight percent then becomes our baseline in that if our model made the same simplistic prediction of non-default for every customer, it would be right in 77.8% of the cases. Our model then needs to predict higher than 77.8%. Our highest model predicts 4.5% higher than the baseline.

Whereas we prefer to see prediction accuracy percentages in the ninety percent or above range, 82.33% still offers a high degree of confidence.

**Which attributes in the data can we deem to be statistically significant to the problem at hand?**

Our model indicated that the most significant attribute from the given dataset is the Pay 0 (i.e. the most recent repayment status). This is followed by (in order of importance) by the Bill Amount 1 (i.e. the most recent bill amount), Limit Balance, Age, Pay Amount 3, Pay Amount 1, Pay Amount 2.

**Recommendations**

After completing this project, we recommend …

1. Credit One implement our Random Forest model to predict potential customers who may default
2. When a potential customer has been predicted as one to Default, that Credit One pursue other avenues of examining the individual to attempt to determine their credit worthiness
3. If possible, Credit One should collect more demographic and financial history data on their potential customers. With this additional data, a more detailed model with even higher accuracy is possible.