CreditOne

Kelly Booth

Customer Default Analysis

Course 5 Task 3

# **Agenda**

* Overview
* Initial Exploration
* Machine Learning Predictions
* Follow up Exploration
* Questions Answered
* Future Recommendations

# **Overview**

Problem:

Over the past year or so Credit One has seen an increase in the number of customers who have defaulted on loans they have secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away.

Questions to Investigate:

1. How can we ensure that customers can/will pay their loans?
2. Can we approve customers with high certainty?
3. Which attributes in the data can we deem to be statistically significant to the problem at hand?
4. What concrete information can we derive from the data we have?
5. What proven methods can we use to uncover more information and why?

# **Initial exploration**

30,000 records were analyzed from the initial dataset. Preprocessing the data showed that many records had a late payment status during months when there was <= zero balance. These were deemed erroneous and removed to leave 27,963 records remaining.

The exploratory data analysis focused on the following features with key takeaways:

* Gender
  + Women outnumber men 3 to 2
  + Men default at a higher rate of 23.5% to 20.2%
* Age
  + Default rates
    - (18-24) – 27.0%
    - (25-34) – 19.8%
    - (35-44) – 21.3%
    - (45-54) – 23.4%
    - (55+) – 24.4

* Education
  + Default rates – highschool 24.6% > university 23.4% > graduates 18.2%
* Marital Status
  + Default rates – divorced 25.5% > singles 22.9% > married 20.4%

View the full [Exploratory Data Analysis](https://github.com/kbooth15/Course-5-Python/blob/master/Course%205%20Task%202/Exploratory%20Data%20Analysis%20Credit%20One.ipynb) with additional analysis and graphics.

# **Machine learning Predictions**

When using machine learning to predict whether a customer will default the following features were found to be the most important:

* 3 most recent payment statuses
* 3 most recent payment amounts
* Credit Limit

The following models were trained and tested to be able to predict when customers will default.

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Kappa |
| Random Forest | 82.8% | 38.5% |
| Supported Vector | 78.0% | 01.5% |
| K Nearest Neighbor | 79.0% | 07.9% |
| Gradient Boosting | 82.6% | 39.0% |

All four models came in right around 80% accuracy, but Gradient Boosted came with slightly the highest accuracy at 82.6%. For Kappa scores, SVC and KNN performed very poorly while Gradient Boosted came in at slightly the highest again right at 39%. Thus, Gradient Boosted Classifier was chosen.

# **Follow up Data Exploration**

After finding out that credit limit was an important factor when predicting customer defaults, we decided to go back to see if we could find any noticeable patterns in the data that we missed during the initial EDA. We broke the limit balance into the following bins and measured average default rates:

|  |  |
| --- | --- |
| Credit Limit | Default Rate |
| 0-50k | 36.0% |
| 50-100k | 25.8% |
| 100-150k | 23.1% |
| 150-200k | 17.1% |
| 200-300k | 15.2% |
| 300-400k | 12.9% |
| 400-500k | 12.2% |
| 500k+ | 10.1% |

There is an obvious pattern that customers with lower credit limits are more likely to default. Specifically, customers with limits less than 150k have a 23% chance or higher to default. We believe this is a clear area to target to improve our overall scoring.

# **Questions Answered**

1. How can we ensure that customers can/will pay their loans?

There is no full proof way to ensure that customers will pay, but we have ways to better predict if they will.

1. Can we approve customers with high certainty?

Yes. Based off 3 most recent months of credit payment history and limit balance we can predict customer defaults within an 82% accuracy. With more fine tuning this could be improved as well.

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

Customer’s recent payment history is clearly the most significant determinant if they are going to default. Followed by how big their credit limit is. Other attributes should not be entirely dismissed though.

1. What concrete information can we derive from the data we have?

Our approval process obviously gets more lenient when the limit amount gets lower. We need to reassess how much of a credit limit we approve especially for smaller amounts.

1. What proven methods can we use to uncover more information and why?

Continue to look for patterns in the data. This data set is from a six month period in 2005. We could use an updated dataset and re run this analysis and tune it even more to see if the patterns hold up or if any new patterns show up.

# **future Recommendations**

For all credit scoring we need to always be checking most recent payment statuses and amounts paid. These were the most significant metrics for predicting defaults. Then after looking at those if there is further question as to how to score a customer we can look at other metrics such as age/gender/education/marital status. Those should not be forgotten completely but used as secondary analysis tools to recent payment history.

Next, when approving smaller amounts of credit, we need to be stricter. Either by approving smaller amounts until the customer has shown they can pay or by denying some of the customers that we previously would have approved based on the key metrics and data patterns we have discovered.