Efficient Revenue Recovery *Using* Machine Learning

By: Bram Tunggala

Collection agencies spends millions of dollars sending letters and making phone calls in order to potentially receive payment of some sort from the debtor.

Most of these companies are blindly sending letters and making phone calls to make an effort to retain some form of payment.

Can we use machine learning to segment and rank accounts by likelihood of payment using historical successes?

Methodology - OSEMN

Obtain

Identify the dataset(s) to use and extract the data into usable format (.csv, json,xml,etc.)

Scrub

Cleaning the data, delete, and/or fill missing values

Examine the data and understand every feature,, identify errors, missing values, and corrupt records

Explore

Find patterns by using visualizations and charts

Extract features

Use statistics to identify significant variables

Model

Use predictive tools to enhance decision making

In-depth analytics using machine learning

Evaluate and refine model

Interpret

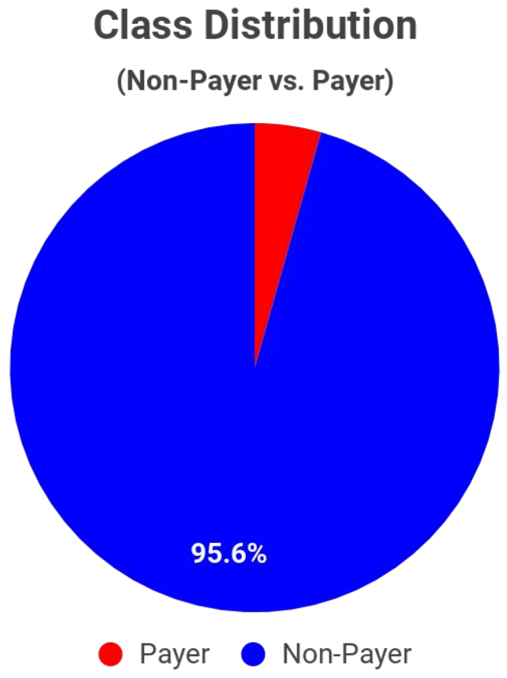
Storytelling through data

Identify insights

Visualize findings

Our Data

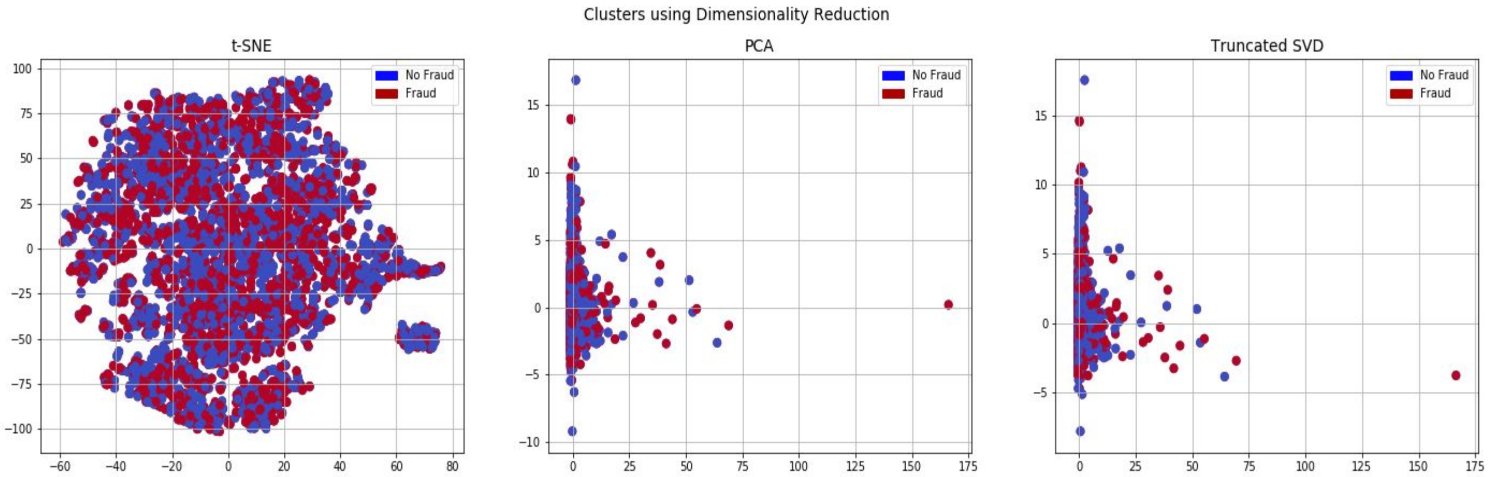
- DB_Accounts_2012-2015.txt: contains account numbers and account specifics
- DB_Splits.txt: contains payment information
- DB_Entities.txt: contains entity address information
- DB_Purchases.txt: contains account balances purchases and descriptive portfolio information
- uszips.csv: contains zip code based economic data
- 12+ million records & 46 columns



Model - Part I

We used three different algorithms in attempt to identify different clusters or our Classes i.e. **“Non-Payer” and “Payer”**

Results: The algorithms failed to accurately cluster the classes.



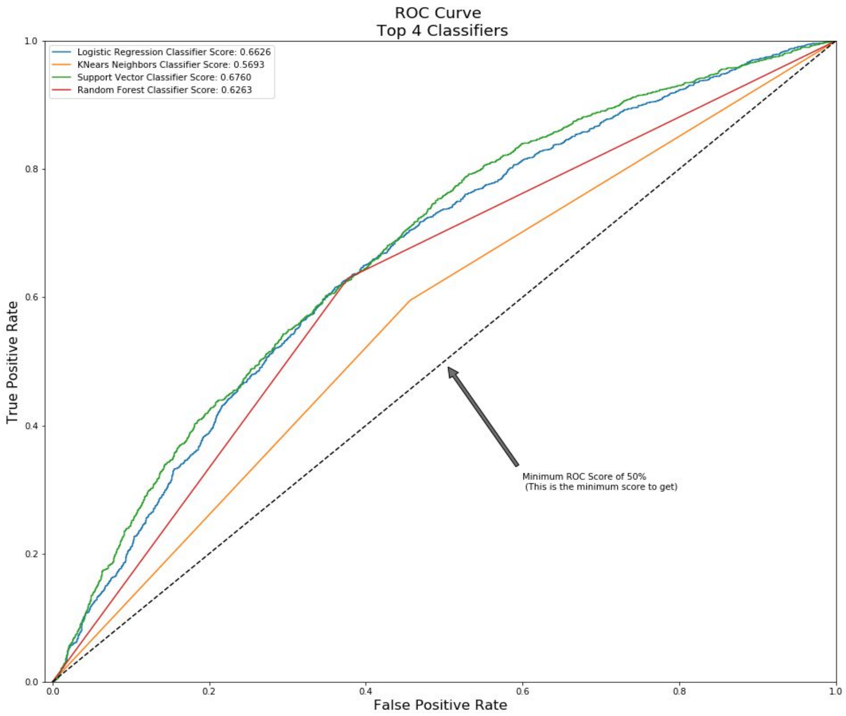
Model - Part II

We used four types of classification algorithms:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine (SVM)
- Random Forest

Results:

- SVM classifier has the best score of **67.5%** which means it identified payers from the non-payers better than the other models



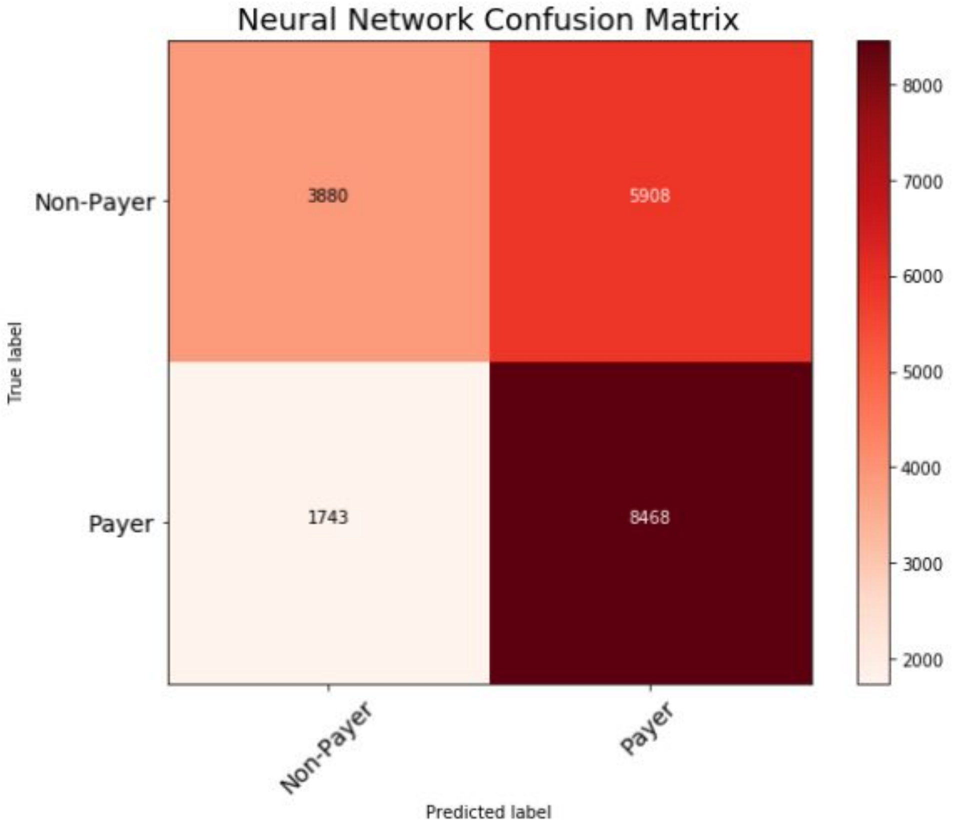
Model - Part III

Simple Neural Networks:

- Group of algorithms that certify underlying relationships in a set of data
- Consist of:
  - Input
  - Hidden layer(s) w/ nodes
  - Output

Results:

- Our neural network identified class "Payer" better than our other models … but, at the cost of increased **miss identification** of our "Non-payers"



Summary

Our SVM classifier model was the best at identifying our classes.

We are not satisfied with the results, however ...

… we extracted important features which our algorithms used to classify our classes, and created a **scoring system** to **prioritize** and **segment** our accounts.

Scoring System

We used our models' feature importance to **score** our feature values, then we categorized into **Grades** by totaling the scores received per account.

features	score
Unemply_rate_range_3-4.9%	10.00
Education_college_5-10%	7.58
Labor_force_part_range_65-67.9%	6.28
Labor_force_part_range_62-64.9%	5.98
Unemply_rate_range_15-19.9%	5.01
Unemply_rate_range_5-6.9%	4.87
Income_level_35k_to_50k	4.02
Income_level_50k_to_75k	3.15
Education_college_10-15%	3.06
Labor_force_part_range_53-55.9%	3.00
Age_range_65-69	2.57

Account_ID	total_score	Grade
10271476	4.51	Dd
10271477	16.71	Aaa
10271478	12.43	Bb
10271479	15.00	Aaa
10271480	14.89	B
10271481	13.01	B
10271482	16.57	Aaa
10271483	12.60	Bb
10271484	4.34	Dd
10271485	16.51	Aaa

Grade	Unpaid_Count	Total_Paid_Count	Total_Accounts	Percent_Paid
A	28245	2014	30259	6.66
Aa	531296	37832	569128	6.65
Aaa	2150150	120472	2270622	5.31
B	1679590	78254	1757844	4.45
Bb	1012650	43971	1056621	4.16
Bbb	2034583	89647	2124230	4.22
C	1982412	84105	2066517	4.07
Cc	1435512	53757	1489269	3.61
D	558840	19219	578059	3.32
Dd	453372	17307	470679	3.68
F	290731	9958	300689	3.31

Segment Accounts

Grade	Age_range	Income_level	Unpaid_Count	Total_Paid_Count	Total_Accounts	Percent_Paid
Aaa	18-21	Over_150k	2	1	3	33.33
Aaa	25-29	Over_150k	45	17	62	27.42
Aaa	45-49	Over_150k	127	34	161	21.12
Aaa	25-29	100k_to_150k	912	202	1114	18.13
Aaa	55-59	100k_to_150k	2470	533	3003	17.75

Grade	Age_range	Income_level	Unpaid_Count	Total_Paid_Count	Total_Accounts	Percent_Paid
Cc	75-79	Over_150k	1	1	2	50.00
Cc	40-44	Over_150k	4	2	6	33.33
Cc	75-79	Under_20k	40	9	49	18.37
Cc	70-74	Under_20k	68	15	83	18.07
Cc	22-24	20k_to_25k	5	1	6	16.67

Grade	Age_range	Income_level	Unpaid_Count	Total_Paid_Count	Total_Accounts	Percent_Paid
F	25-29	Over_150k	12	4	16	25.00
F	45-49	Over_150k	17	4	21	19.05
F	25-29	100k_to_150k	211	40	251	15.94
F	30-34	Over_150k	11	2	13	15.38
F	18-21	100k_to_150k	6	1	7	14.29

Next Steps

- Scrape zip code based consumer behavioral data (if possible) and combine with our current features.

- Use sensitive information i.e. ss#, gender, sex and race.

- We would like to create a pipeline from receiving our data, OSEMN process (clean,modeling, ect.), and create a user friendly dashboard that collection agencies can utilize.

- We also would like to compare our scoring system i.e. "likelihood" of payment vs. other scoring systems in the marketplace.

THANK YOU!