



An Analytics Course Reflection Project

July 11th, 2018



Introduction

DirectPay is a debt collection agency based out of the Netherlands. DirectPay buys portfolios of delinquent invoices or loans from companies like telcos, utilities, and online stores.

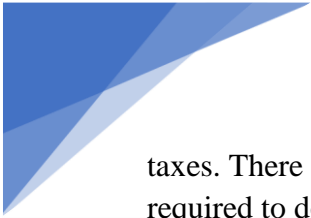
Since debt can be a complex issue, it is hard to read the consumer behavior through excel spreadsheets. Colin Nugteren, the Chief Analytics Officer for DirectPay, used to have a lot of spreadsheets. He helps DirectPay to decide on what debt to buy and how best to collect it. The company used SAS software that helped them visualize current information, explore relationships, provide predictions and display it all on a mobile device when they visit their clients.

Analytics Models Used

Since DirectPay buys portfolios of delinquent invoices, each portfolio could include millions of delinquent invoices or loans. DirectPay was trying to understand the likelihood that it can collect the money owed by the individual debtors to set a reasonable bid price. Before the company uses any analytical models, it is vital to investigate the data that has been collected by the company about its customers. In order to do this, the company will have to go through the vast amount of data from its Data Warehouse system and had to carefully investigate the below factors:

1. Types of portfolios: Since DirectPay works with business customers from various industries, there should be a categorization for portfolio types. Example includes: Retail, Utilities etc. This will be a categorical variable.
2. Debt owed by each customer: This should indicate the total amount of debt owed by each customer. This will be a quantitative variable.
3. How long has the customer not paid their debt: This will be a quantitative variable.
4. Customer age: Often age of the customer can reveal a lot about customer's ability to pay the debt. For example, a high school or a college student might typically not be financially independent or may be dependent on a parent to pay bills. A college student could also be on a student loan which might hinder him/her from paying other types of bills. This will be a quantitative variable.
5. Customer employment status: Is the customer employed or unemployed. This will be a categorical variable.
6. Customer Income: What is the annual income earned by the customer.
7. Are there any other debts owed by the same customer? And if so how much? A college student might have student loans, utility bills, online purchases bills etc. This will be a combination of categorical and quantitative variable.
8. Customer's credit score: This will be a quantitative variable.

After extracting information on the above factors, the company should then see if there are any missing data and if there is a pattern that follows the missing data. For example, a customer who is earning a lot of income might not report information about their income to sometimes evade



taxes. There could be scenarios where the customer will not answer questions unless they are required to do so. For example, when I created my amazon account, I tried to give as little information as possible and avoided questions like date of birth, country of birth etc. The company will have to reach out to its portfolio owner to get more information on data that is missing. If there are any missing data for the below factors, below are possible solutions that could be used to estimate a value for missing data:

1. Income (Numeric): Use regression and perturbation
2. Credit score (Numeric): Can use mean, median, or mode

After identifying patterns in the missing data, the company could then use support vector machine learning model to classify its debtors as high risk, medium risk, and low risk.


1. High risk: customers with high debts, credit score, duration of debt etc.
2. Medium risk: customers with medium debts, credit score, duration of debt etc.
3. Low risk: customers with low debts, credit score, duration of debt etc.

Clustering model can also be used to create same classification by taking past data on debt owed by each customer and their credit scores. For example, a person with a higher credit score and higher debt might get classified into high risk category. At a larger scale, the company can also categorize its portfolio based on risk with this model. Both of the above models should be tested with training, validation, and test data set. As we have studied in this course, correlation does not always mean causation. After categorizing their customers, they could use regression to estimate whether there is in fact a correlation between risk level and the likelihood of someone paying their debts.

Additionally, DirectPay recently launched a new independent company called DebtScan to find cars that are subject to repossession. A car can only be repossessed in cases in which a court grants a verdict to enforce debt collection. DebtScan drivers use cars equipped with cameras that capture license plate data. The data is matched against cars owned by people behind on their payments, and the nearest law enforcement office is notified of a car's location.

DebtScan should have acquired past data specifically about cars that are subjected to debt collection. In order to identify which cars from their Data Warehouse is subjected to debt collection, they should have analyzed the below factors:

1. Price of the car
2. Amount of debt owed on that car
3. How long has the customer not paid their car loan
4. Is the car still being utilized?
5. Customer location
6. Customer phone number



As mentioned before they could use a simple support vector machine learning model to create classification based on risk. For example, high, medium, or low risk. Upon classifying their debt holder, they might have approached the court to get a verdict on collecting debt on customers who belong to high or medium risk categories. After getting permission, they would have created a cluster model to cluster customers based on location. Before driving to the customer location, which requires resources, they could have used an automated simulation to give their customers a call to turn in their cars. If the customer does not respond, the drivers go to the customer location in the clusters. In each cluster, they start approaching customers with the highest risk first and then medium risk.