Open IIT Data Analytics

TEAm NAme:

The ten commandments

Customer Lifetime

Value Prediction

# Introduction

Auto insurance is a policy purchased by vehicle owners to mitigate costs associated with getting into an auto accident. Instead of paying out of pocket for auto accidents, people pay annual premiums to an auto insurance company; the company then pays all or most of the costs associated with an auto accident or other vehicle damage.

In exchange for paying a premium, the insurance company agrees to pay your losses as outlined in your policy. Coverages include:

* Property – damage to or theft of your car
* Liability – legal responsibility to others for bodily injury or property damage
* Medical – costs of treating injuries, rehabilitation and sometimes lost wages and funeral expenses

Customer Lifetime Value (CLV) is the total revenue that a customer generates for the company. As length of customer relationships may vary, we state CLV as a periodic value.

In the given problem statement, we are given the data of different customers of an auto insurance company containing parameters such as their income, monthly premium, claims, number of policies etc. and we have to predict CLV accordingly. We also need to find out which kind of customers are the most beneficial which would thereby help the company’s marketing, advertisement and sales team to focus their endeavors towards the more beneficial demographic. If the cost of acquiring certain customers exceeds their CLV it is detrimental to the company’s profits and thus these customers mustn’t be provided insurance or their premium must be increased such that CLV becomes greater than cost.

# Data Preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format

## Data cleaning

First, we looked for any missing data, but the data was complete, there were no missing values

## Data Transformation

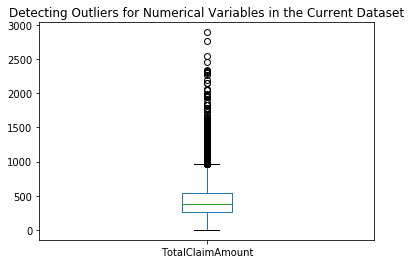
* Anomaly Detection: We analyzed the data manually to find some logical discrepancies in it. We found two major flaws and to improve our prediction we removed those from our data

1. The customers who had only 1 policy and their months since last claim was higher than the months since policy inception were removed because it is impossible to claim a policy before its inception
2. The customers who were unemployed, had 0 income and had customer lifetime value greater than the upper threshold of CLV (as defined below) were removed because a general trend and a positive correlation between income and CLV show that these data points are essentially noise and hinder accurate prediction

Interquartile range= 3rd quartile – 1st quartile

Upper Threshold of CLV= 3rd Quartile + 1.5\*Interquartile range

* Outlier Removal: Outliers are data points that don’t belong to a certain population. It is an abnormal observation that lies far away from other values. An outlier is an observation that diverges from otherwise well-structured data. We have found outliers using boxplots of various features and observed the following features had outliers that needed to be removed to improve our prediction



Some other features such as income and monthly auto premium also had similar outliers but as those weren’t a small number and people with higher income and monthly premium are the best customers of the company, we chose not to remove them. On the other hand, claim amount can be absurd due to very rare occurrences causing high claim amount so we removed these outliers.

* Scaling: We used the min-max scaler to scale the data in the range of 0 to 1 to transform the data such that it follows normal distribution

