**Optimization Final Project Report**

**By: Souraj Krishnan, Kanish Shah, Minal Chandra, Chenxiao Xu**

**Topic:**

Towards an optimum prediction of customer response for the auto insurance industry

**Motivation:**

* In present times, predictive analysis has become a vital aspect in business decision making of major companies in various industries such as the insurance industry.
* Insurance companies are looking for insights into the reason for low customer response rate, which is affected by various factors.
* Predicting the response of a customer based on features such as Age, Gender, Profession, Salary, etc. can help insurance companies optimize their use of time by focusing on only potential customers.

**Problem Statement:**

* The Goal is to predict the response of a customer as “yes” or “no” for a promotional product from an auto insurance company, for customer data derived from the IBM Watson dataset.
* This experiment can be broadly classified as a classification task.

**Dataset:**

The dataset that we worked on for this project consisted of 9134 rows and 24 columns from an auto insurance company. The 24 columns were the variables that we dealt with to solve for our goal. The variables are mentioned below:

1. Customer
2. State
3. Customer Lifetime Value
4. Response
5. Coverage
6. Education
7. Effective to Date
8. Employment Status
9. Gender
10. Income
11. Location Code
12. Marital Status
13. Monthly Premium Auto
14. Months Since Last Claim
15. Months Since Policy Inception
16. Number of Open Complaints
17. Number of Policies
18. Policy Type
19. Policy
20. Renew Offer Type
21. Sales Channel
22. Total Claim Amount
23. Vehicle Class
24. Vehicle Size

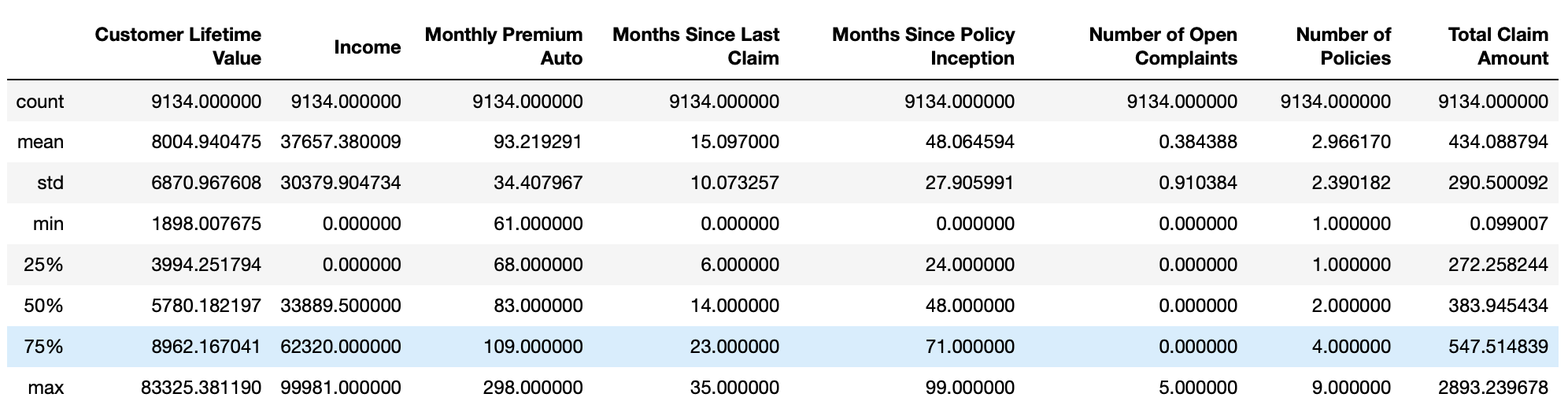
The variables Customer and Effective to Date were not used while preparing our data to run our models as they were unique identifiers which were not necessary to us. The rests of the variables were required and were either continuous or categorical. The dataset was split into 70% of training set and30% of test set. The target response variable for our models was the Customer Response which would either achieve a positive “Yes” or a negative “No” response.

**Exploratory Data Analysis:**

Below is a list of the continuous and the categorical variables from our dataset.

|  |  |
| --- | --- |
| **Continuous** | **Categorical** |
| Customer Lifetime Value  Income  Monthly Premium Auto  Months Since Last Claim  Total Claim Amount  Number of Open Complaints  Number of Policies  Months Since Policy Inception | State  Response  Coverage  Education  Vehicle Size  Gender  Location Code  Marital Status  Policy Type  Policy  Renew Offer Type  Sales Channel  Vehicle Class  Employment Status |

We ran a descriptive analysis of the continuous variables to get a better understanding of the types of values we were dealing with to understand their mean, standard deviation and range.



After this, we ran a Z score calculation to identify any outliers that were present in any of our continuous variables. Once identified, we quickly removed them from out data set. Also any missing values were replaced by an average measure as we thought getting rid of the entire row could lead to a bias in our models due to under fitting.

The categorical variables were converted to numerical values using One Hot Encoder to prepare the data for our machine learning models to run on. The table below shows the list of categories and their identification.

|  |  |
| --- | --- |
| State | Washington  Nevada  Arizona  California  Oregon |
| Response | No  Yes |
| Coverage | Basic  Extended  Premium |
| Education | High School or Below  College  Bachelor  Master  Doctor |
| Gender | Female  Male |
| Location Code | Suburban  Urban  Rural |
| Marital Status | Married  Single  Divorced |
| Policy Type | Personal Auto  Corporate Auto  Special Auto |
| Policy | Personal L1  Personal L2  Personal L3  Corporate L1  Corporate L2  Corporate L3  Special L1  Special L2  Special L3 |
| Renew Offer Type | Offer 1  Offer 2  Offer 3  Offer 4 |
| Employment Status | Employed  Unemployed  Disabled  Medical Leave  Retired |
| Sales Channel | Agent  Call Center  Web  Branch |
| Vehicle Class | Two-Door car  Four-Door car  SUV  Luxury SUV  Sports car  Luxury Car |
| Vehicle Size | Small  Medsize  Large |

We created a few pair plots to explore how each variable reacted to another variable. A sample size of plots is shown below. These plots helped us get an idea of which of the variables could be dependent on each other.

A picture containing building, window, room

Description automatically generated

We also created a few graphs to find out how the independent variables reacted to our target response variable. An example of this is shown below:

A screenshot of a cell phone

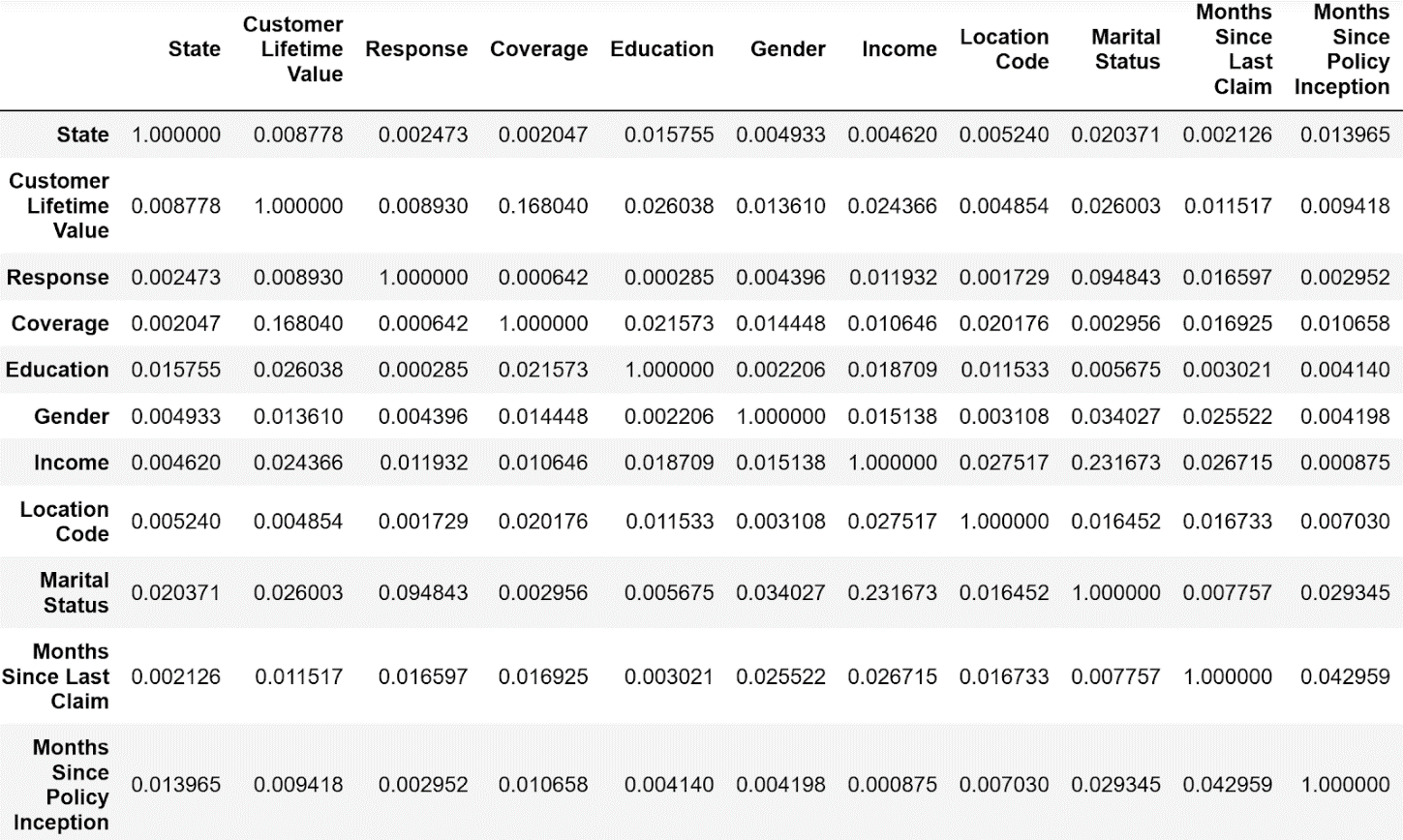
Description automatically generated

This graph shows how each gender reacted to the target response as it shows the percentage of male and female that either said “Yes” or “No”. This graph can also later be used to compare our predicted results with the actual result.

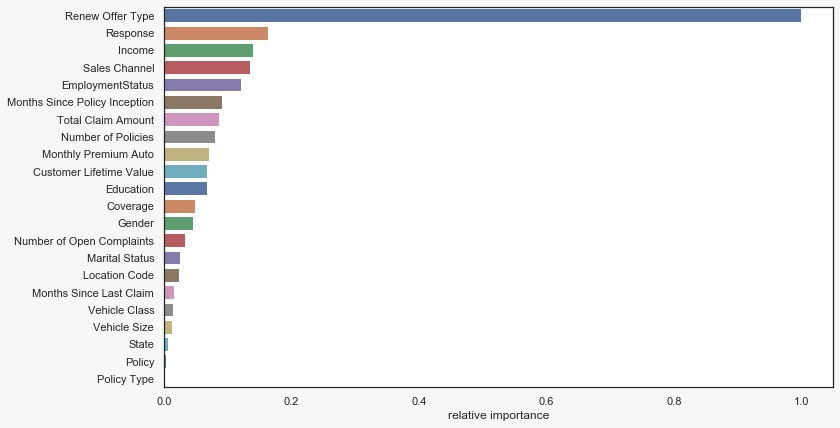
**Feature Selection:**

The next step is the feature selection since we have 22 columns which are relatively high, and we need to drop some of it to avoid noise.

As most of the attributions in the table are categorical, we used label encoder to transform all categorical value to numerical value with numbers scaling from 1 to 10. And then in order to make the correlation matrix and the heat map, we must scandalize the data to have the values spreading from 0 to 1 so that they are close enough to do the research, the below shows the data after the label encoder and the standardization.



Next, we simply calculate the variance of each of the attributions with the outcome to come up with a chart of feature importance as it is showed below:



Other than the top one feature renew offer type, most of the features is only having an importance between 0 and 0.2 which made it hard to select. We dropped all features having an importance lower than 0.1 from our dataset at last to help avoid noise and then we go proceeded to model selection.

**Machine Learning Models:**

1. Naïve Bayes

As it is a classification problem with a binary outcome, our team came up with Naïve Bayes model which is good at dealing with large data size and is easy to use.

Description: A picture containing bird

Description automatically generated

The logic behind using this model is that we need to find out the probability of each customer picking up the phone and giving a feedback to our promotion. The accuracy turns out to be approximately 0.85 which is high. The precision matrix below shows that 0.0 stands for the customer not interested in the products that we are promoting. At that point, the precision is 0.86 which tells us the model is doing well on helping us filtering out customers who could waste the sellers teams’ time. However, the precision of 1.0 which stands for customers who are willing to participate in the product we promote and will possibly give feedback to us is quite low (0.43). We cannot simply stop and pick this as our final model because the model could filter out several potential customers who are valuable to the sales team. So we went onto the next model for prediction.



Description: A picture containing clock, meter, drawing

Description automatically generated

1. Random Forest Classifier

The next model that we chose to use is Random Forest Classifier. The random forest is a model made up of many decision trees. This model uses two key concepts that give it the name random rather than simply averaging the prediction of trees:

1. Random sampling of training data points for trees
2. Random subsets of features taken into account when nodes are split

The random forest incorporates hundreds or thousands of decision trees, trains each on a slightly different set of observations, splitting nodes in each tree with a limited number of characteristics. The random forest's final predictions are made by an average of each tree's predictions.

Description: A picture containing text, map

Description automatically generated

Description: A screenshot of a cell phone

Description automatically generated

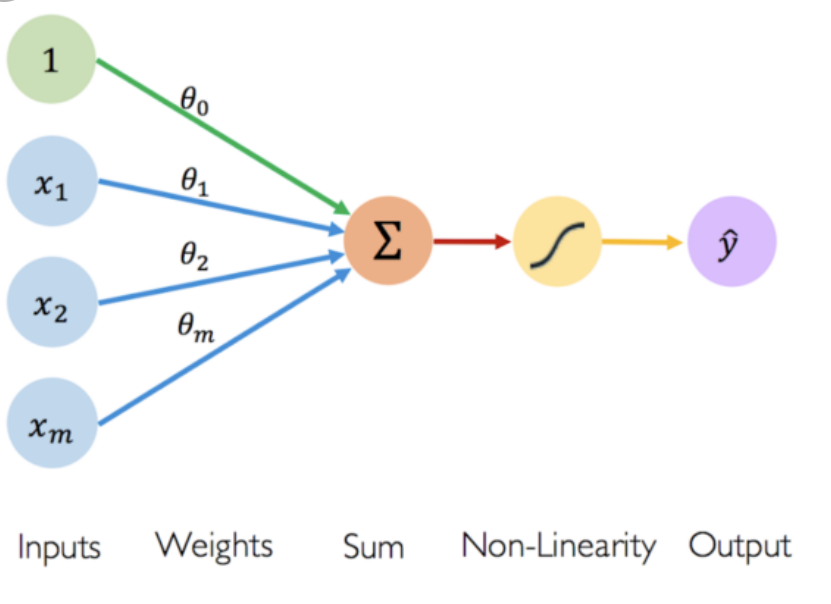
Description: A picture containing device, meter

Description automatically generated

An accuracy of 80.79% is achieved through implementing the Random Forest Machine learning Model which was not as good as the accuracy that we achieved for Naïve Bayes and due to this we decided on trying out another model for prediction.

1. Artificial Neural Network:

* ANNs are a very rough model of how the human brain is structured. An Artificial Neural network is a supervised machine learning model that is used for the task of classification.
* The basic building block of an ANN is a neuron, and each neuron consists of a summation function and an activation function – the activation function is usually a sigmoid function.

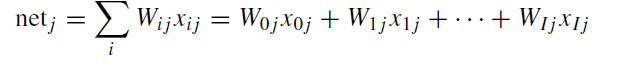


How does it work?

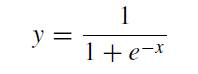
* An ANN has 3 layers:

1. Input Layer
2. Hidden Layer
3. Output Layer

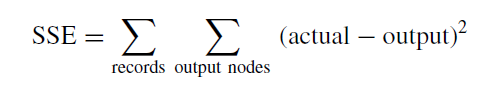
* Each layer consists of neurons, and every neuron is supplied with the incoming input from another neuron.
* Every path between two neurons consists of a unique weight which is set to 0 or 1 randomly in the beginning.
* Inside each neuron the sum of the product of input and the respective weight is calculated and passed to the activation function.



* The activation function is a sigmoid function that returns a value which is compared to a randomly generated threshold.

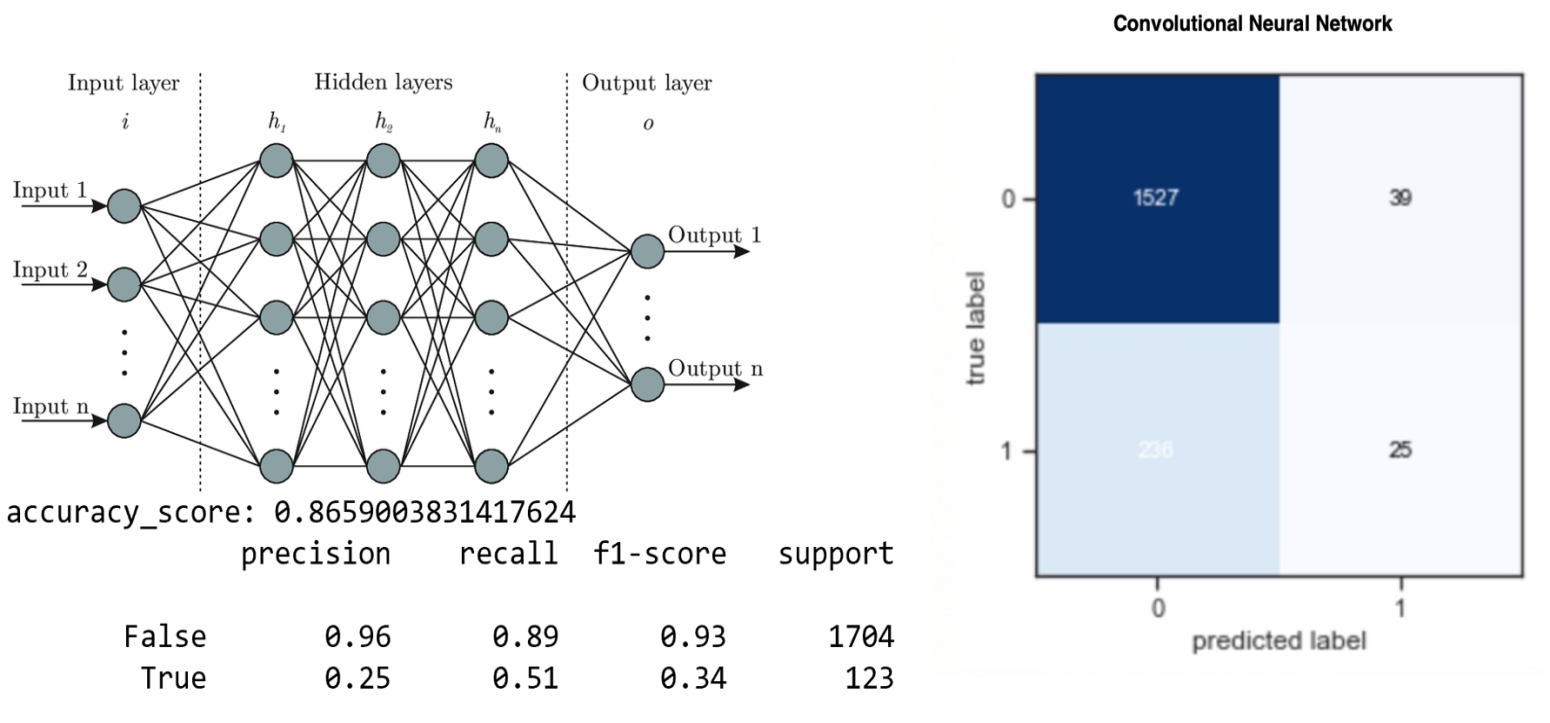


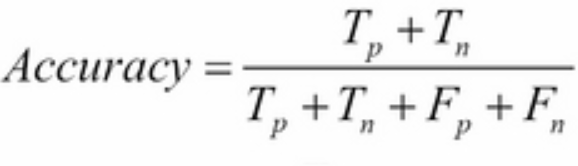
* If the returned value is greater than the threshold, the neuron is triggered and sends its output to the next layer.
* If the predicted output does not match the actual output, using Back-propagation gradient method the sum of squared errors (SSE) is found to be:

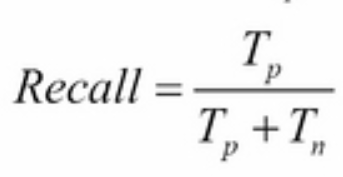


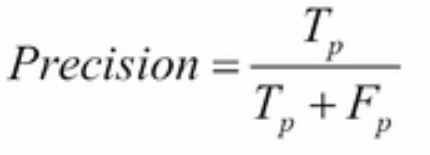
This is then used as a benchmark to trace back through the series of neurons and reassign new values for the weights.

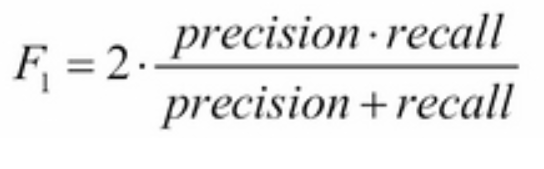
* The Model does this repeatedly during training till it achieves ideal values for the weights.
* Accuracy of 86.59% was achieved using the Artificial Neural Network Model











**Conclusion:**

Artificial Neural Network turned out to be the best model of the three classifiers as it had 86.59% accuracy for the task of classification. This model can now be applied to predict the outcome for an auto insurance company when it promotes a product as it would know whom to specifically promote the product to as from the machine learning model it has learnt that only a specific customer would have a positive response to a certain product due to a certain variables that can be called as the best predictors.

In the future it would be interesting to also use clustering as it could help to perfect customer segmentation for the marketing of personalized policies/products.