

INSURANCE CROSS-SELL PREDICTION



**Problem statement:**

When a Health Insurance company wants to add a Vehicle policy business to its portfolio, an analysis should be made to understand how well the business would perform in real time. This can be done by collecting some data from existing customers and understanding if they would show interest in buying their new product. Our project aims to provide the best model for predicting the likelihood of a customer opting for the Vehicle policy.

**Data sources:**

Fetched data from Kaggle. The dataset consists of approximately 380,000 rows and 12 columns of data related to existing customers and their vehicles. Initially, used whole data for the project but later, due to the system’s limitations, had to use only 50,000 rows.

The data has 11 input columns and 1 output column as specified below.

**Inputs:**

* Gender – If Male it’s stored as 1 and for Females, it is 0
* Previously insured – If the customer is already insured, Previously insured is stored as 1. Otherwise, it’s 0
* Vehicle Age – Age of the Vehicle
* Vehicle Damage – If the vehicle is damaged previously, it’s stored as 1. If the vehicle is never damaged, it’s stored as a 0
* Driving License – If the customer has a driving license, it’s stored as 1. Otherwise, it’s stored as a 0
* Annual Premium – Premium that customer has to pay for the new policy
* Vintage – Number of days the customer has been associated with the company
* Region code – Code of the region that customer lives in
* Age – The age of the customer
* Policy Sales Channel – Channel the customer was approached. Examples – email, telephone
* ID – unique ID of the customer

**Outputs:**

* Response – If the customer is interested in taking a new vehicle Policy, the response is stored as 1. Zero otherwise

The dataset has categorical columns such as Gender, Previously Insured, Vehicle Age, Region Code, Policy Channels, and Vehicle Damage. The rest are numerical columns.

**Dataset link:** <https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction>

**Data Preparation:**

As there are no missing values and no duplicate data, no cleaning has been done. Two different encoding techniques as **Ordinal encoding** and **One hot encoding** are usedto visualize and for modeling respectively. Initially, Ordinal encoding has been used for data prediction, but we later moved to one hot encoding just to understand how it would work.

Scaling techniques such as **Minmax scaling** and **Standard scaling** have been done on training and testing data and comparisons were made with scaled vs unscaled data to see which model would give us better results.

Similarly, Sampling methods such as **Random Oversampled**, **Random Undersampled,** and **SMOTE** were used on Minmax data and Standard scaled data to see the change in performance and efficiency of different models.

Initially, used outputs from different sampling methods as inputs for the model, but later reduced this process by using **Pipeline**.

The data is split in an 80-20 ratio for training and testing.

Have also tried to use only variables that have a correlation greater than 0.1 and less than -0.1 to see how the model would behave but later decided not to as it would only leave us with three parameters out of 10.

**Data Visualization:**

The Correlation heatmap has been plotted for different variables and the result is as below

Treemap chart

Description automatically generated

Variables such as Age, Previously Insured, Vehicle Age and Vehicle Damage showed very high correlation compared to other variables.

**Age vs Response:**

**Chart, box and whisker chart

Description automatically generated**

Customers who are above 35 years of age are more likely to choose new vehicle insurance while customers below that age are not interested at all.

**Previously Insured vs Response:**

**Shape, rectangle

Description automatically generated**

Customers who do not have insurance already are more likely to choose the new vehicle insurance policy.

**Region code vs Response:**

Chart, bar chart

Description automatically generated

People living in Region 28 showed interest in considering a new vehicle policy

**Policy Sales Channel vs Response:**

**Chart, bar chart

Description automatically generated**

Channel\_26.0 has received more positive responses from customers when compared to other channels

**Driving license vs response:**

****

Only the customers having a driving license showed interest in choosing vehicle insurance. There is a negligible count of people who does not have a driving license opting for the new vehicle policy.

**Vehicle damage vs response:**

A picture containing graphical user interface

Description automatically generated

Customer, whose vehicle is damaged previously is more likely to choose the vehicle insurance

**Gender vs response:**

**Chart, bar chart

Description automatically generated**

Males show slightly more interest in vehicle insurance compared to the females.

**Vintage vs response:**

**Chart, box and whisker chart

Description automatically generated**

Customers who are associated with the company for more than 100 days are more likely to choose the vehicles insurance

**Vehicle age vs response:**

Chart, bar chart

Description automatically generated

**Vehicle Age 1-2 Years = 0, < 1 Year = 1, > 2 Years = 2**

Customers, whose vehicle age is in between 1-2 years are more likely to choose a vehicle insurance

**Modeling:**

Have used three Machine Learning models in this project. They are as below

1. Logistic Regression
2. KNN Classifier
3. Random Forest Classification

The inputs for these models come through different scaling and sampling methods. For example, in Random Forest Classifier, we have 7 different inputs as below

1. Standard scaled Random Undersampled data
2. Standard scaled Random Oversampled data
3. Standard scaled SMOTE’d data
4. Minmax scaled Random Undersampled data
5. Minmax scaled Random Oversampled data
6. Minmax scaled SMOTE’d data
7. Unscaled and Unsampled data

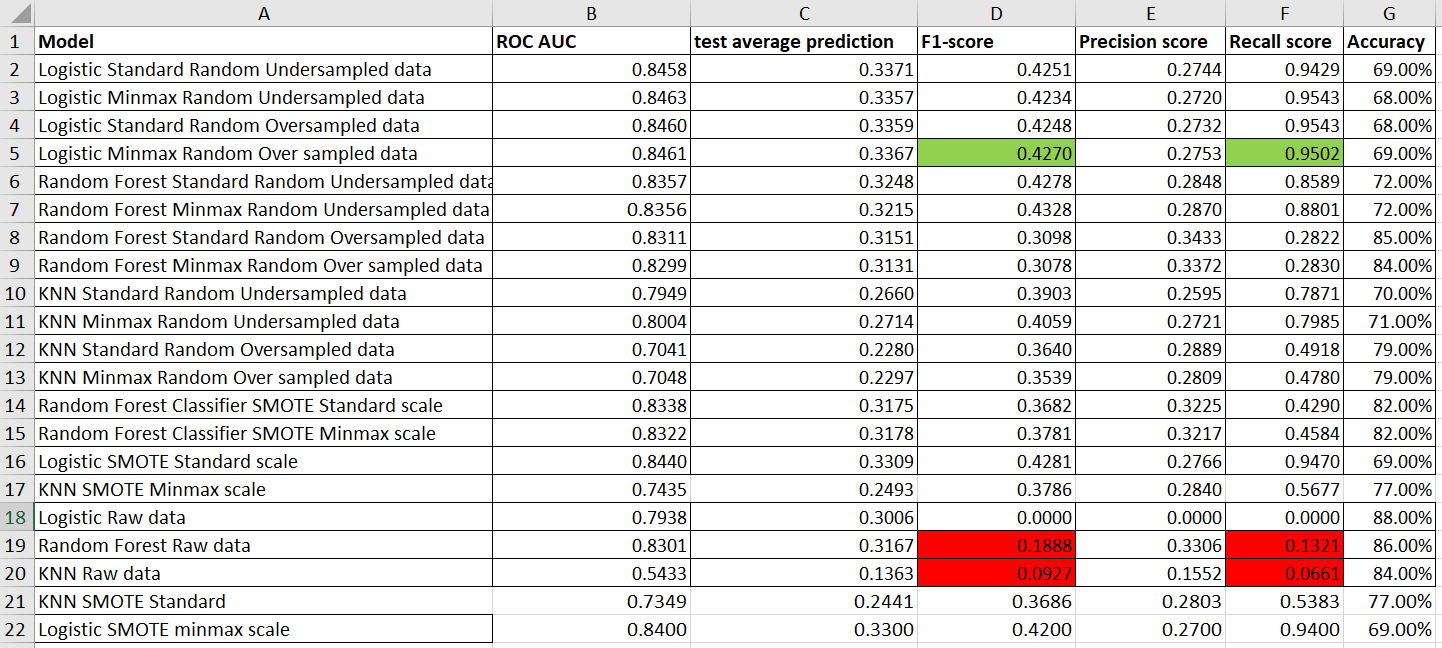
Similarly, there are 7 different inputs for each KNN and Logistic regression which gives 21 different inputs in total.

In order to measure efficiency, calculated the F1 score, Precision score, Recall score, ROC AUC mean and Average prediction mean, Confusion Matrix, and classification report.

Used Jupyter notebook primarily to develop code. However, due to the system’s low processing speed, had to move to Google collab. Later came back to Jupyter as the number of rows is reduced due to excess consumption of RAM in Google collab.

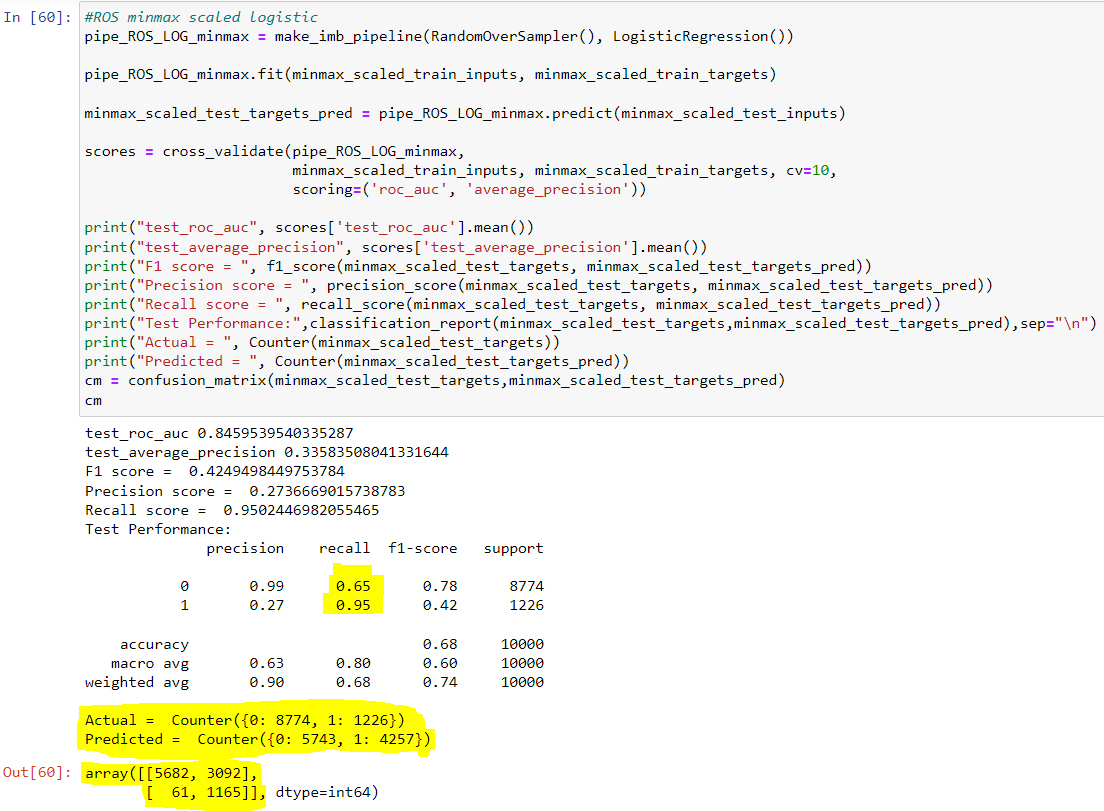
**Evaluation:**

As the data set is imbalanced, have compared F1-score and Recall. And found [this article](https://medium.com/datasciencestory/performance-metrics-for-evaluating-a-model-on-an-imbalanced-data-set-1feeab6c36fe) helpful in determining key metrics that are to be considered when dealing with imbalanced data with negative values as the majority.



**Better performing models:**

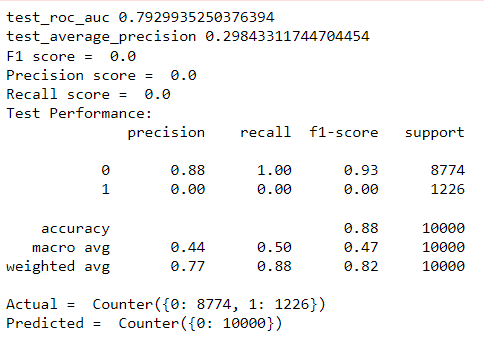
From the derived results, as shown above, have observe a high F1 score and Recall score for Logistic Regression with Minmax scaled and Random Oversampled data. We do observe similar results for Logistic Regression with Standard scaled and Minmax scaled data when data underwent Random Undersampled, Oversampled, and SMOTE sampling methods. Below is the result – Confusion matrix, Recall, and comparison of Actual and Predicted count of 0s and 1s.



**Poor performing models:**

Though the accuracy is comparatively higher for Random Forest Classifier, it gave a very low Recall score. It is evident that Accuracy is not an appropriate metric to be used when we are dealing with highly imbalanced data as the models that took raw data as input gave very high accuracy but when it comes to the Confusion matrix, F1 score, and Recall, they are the least performing models. The outputs of the Confusion matrix as attached below for the model which predicted that there are no 1s.





**Key observation:**

The model should be having the False Negative count as low as possible. The intuition behind this is, if the model predicts a customer who is actually interested in taking the policy as “not interested”, the company is losing a customer. Whereas, if the model predicts a customer who is not interested in the company as “interested”, it’s not an issue for the company. We observe TN count is very low for Logistic Regression with scaled and sampled data.

**Recommendations to Decision-makers:**

1. As data contains more observations with negative responses, though the annual premium the company is offering is not impacting response, customers are not showing interest due to other reasons which cannot be controlled by the company. Due to a lack of demand, the company should reconsider the decision in introducing a new product.
2. Company should come up with better incentives in order to lure customers in opting for their product.
3. From the graph Visualization graphs, it can be observed that customers living in Region-28 and Region-8 show interest in buying their policy when compared with other regions. It would be ideal to start issuing Vehicle Policies in the mentioned regions and then expand toward other regions.
4. It can also be observed from the graphs that when the company approached customers through Channel-26 and 124, they received better responses compared to other Channels. We would recommend the company focus on reaching out to customers via these Channels.
5. From the correlation heatmap, it can be observed that if the customer already has vehicle policy, he is not likely to consider a new policy offered by the company. Hence they can exclude such customers and reach out more to customers who are older in age and customers whose vehicles is damaged in the past.
6. Another observation from heatmap is that the Annual premium offered by the company is not a key factor that’s impacting responses. Hence rates need not be lowered.