Author : Zeeshan Ahmed

Course Title: Health Insurance Cross sell Prediction

Keywords: Machine Learning,Logistic Regression

Abstract:

Often we see people purchasing Health Insurance for themselves and their family. Health Insurance is crucial for betterment of their lives. Vehicle Insurance is also important too but not as much as Health Insurance.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue.

We are required to a dataset of customers who purchased Health Insurance in the previous year.

1. Problem Statement

We are tasked with predicting whether a customer who previously purchased a company's Health Insurance will opt for Vehicle Insurance or not. This is a supervised machine learning classification problem with many independent variables and dependent variables namely Response which comprises responses of customers regarding vehicle insurance.

Our model helps understand the behaviors of customers and build a model around that behavior.

In our Data we've following columns:

1. **ID** : Unique ID for the customer
2. **Gender** : Gender of the customer
3. **Age** : Age of the customer
4. **Driving_License:** 0 : Customer does not have DL, 1 : Customer already has DL
5. **Region_Code** : Unique code for the region of the customer
6. **Previously_Insured:** 1 : Customer already has Vehicle Insurance, 0 :Customer doesn't have Vehicle Insurance
7. **Vehicle_Age** : Age of the Vehicle
8. **Vehicle_Damage 1** : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
9. **Annual_Premium** : The amount customer needs to pay as premium in the year
10. **PolicySalesChannel** : Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
11. **Vintage** : Number of Days, Customer has been associated with the company
12. **Response** : 1 : Customer is interested, 0 : Customer is not interested

2. <u>Steps Involved</u>

1) Exploratory Data Analysis :-

EDA helps us to understand data in a much easier way. It gives us some valuable information that is helpful in model building as well as useful

in outside-model-work. In our data we plot bar plots, pie chart in order to gain insights that might be useful to understand data/people

2) Null Value Treatment :-

It is said that data with more than 40% null values should be dropped as they only drag down the model if used. If treated then won't have much effect either and so getting any column at the beginning and to clear it out is very helpful. In our case when checked for null values we didn't have any.

3) Multicollinearity :-

Extracting correlation heatmap and calculating VIF to remove correlated and multicollinear variables. Implemented resampling to remove the class imbalance of dependent variable values i.e, Response column

4) Conclusions from EDA:-

- There is a great disparity among positive and negative responses from customers.
- Among the positive responses, males were more interested in purchasing Vehicle Insurance.
- PolicyHolders between age groups 27-45 were most interested in vehicle insurance
- Most negative responses are from the age group 23 and 24 years
- The customers who possess driving licenses almost always purchase vehicle insurance.
- People who don't already have a vehicle insurance policy opt in for

vehicle Insurance.

- If the customers' vehicle is damaged, they definitely will buy vehicle insurance as seen in the data.

- There is substantial response from Area code 28 followed by codes 8 and 46.

5) Training the model:

- Assigning the dependent and independent variables

- Splitting the model into train and test sets.

- Transforming data using StandardScaler.

- Fitting logistic regression on train set.

- Getting the predicted dependent variable values from the model.

6) Model performance

A) Confusion Matrix :- The confusion matrix is a table that determines how successful a model is at prediction.

B) Precision/Recall :- Precision is ratio of correct predictions to the overall number of positive predictions: TP/TP+FP. Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

C) Accuracy :- Accuracy is given by the number of correctly classified examples divided by the total number of classified examples.

D) f1 score :- It considers both Precision and Recall of the test to compute score, F-Score is the Harmonic mean of precision and recall. This will tell you how your system is performing.

7) Models Used:

1. Logistic Regression:

   In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination).

   Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

   The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name.

2. Random Forest:

   Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

   For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.However, data characteristics can affect their performance.

3. XGBoost Model:

XgBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

Conclusion

We are finally at the conclusion of our project! Coming from the beginning we did EDA on the dataset and also cleaned the data according to our needs.After that we were able to draw relevant conclusions from the given data and then we trained our model on logistic regression and other models Out of all models used , with the XGBoost classification model we were able to get the F1-score of 0.80.The model which performed poorly was Naive Bayes Classification model with r2-score of 0.73. Given the size of data and the amount of irrelevance in the data , the above score is good.