Health Insurance Cross Sell Prediction

Nikunj Sonule
Data Science Trainee
Almabetter, Bangalore

Abstract:

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee. We were provided with 'Health Insurance Cross Sell Prediction' dataset to perform machine learning task, were to get insights from the independent and dependent variable of a dataset.

Our experiment can help understand what could be the reason for the classification of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the classification problems.

Keywords:machine learning, classification, algorithms

1. Problem Statement

Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company. By analyzing the problems with 'Response' feature whether the customers are eligible to claim the Insurance or not.

The company insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called 'sum assured') to the customer.

We've all record and data with different Age, Driving License, Region Code, Previously Insured, Vehicle Age, Vehicle Damage, Annual Premium, Policy Sales Channel, Vintage, Response (Target variable – 1 and 0). Providing specific data analysis and prediction to done with this data. The main objective is to built a predictive model, which could help in predicting the – whether the customer would be interested in Vehicle insurance. This would help them in providing Insurance to a right specific customer.

- id : Unique ID for the customer

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

2. Introduction

Vehicle insurance every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation to the customer. The major concern is about which customer are truly interested in Insurance. Looking towards various factor/feature affecting the 'Response' (i.e dependent variable). It is important to make Insurance available to customer according to their Annual Premium (the amount they pay to Insurance company).

Generally the algorithms can predict best fit model for the data according to it roc_auc score.As for classification we'll be focusing on roc_auc score. This Insurance data can shows how it varies from every machine learning model. So, generating different roc_auc score from various different model to know which is better.

Our goal is to build a predictive model, which could help company in predicting whether a customer would be interested in Vehicle Insurance is extremely helpful for the company.

3. Customer claim Insurance – Yes or NO

- **Observations from 'Response' feature**

As per the Insurance policy by the company, if the customer is right to provide Insurance or not, is a main important thing. The dataset is highly imbalanced as customer who are interested are 46710 (i.e 12.25% of overall data) and customer who are not interested are 334399 (i.e 87.74% of overall data).The data for Male category is greater as compared to female category. The estimation to claim Insurance for Male

is similar as compared to Female. Customer who are not previously insured are likely to be interested in Insurance. As analysis is very accurate about customer to estimate. They are more likely to interested to claim Insurance.

4. Independent feature factor affecting

Young customer are not much well known to Insurance and are glad to sign up in Insurance. But there is quite variation when we look at different Age groups. The Middle age customer and Old age customer are more likely to claim the Insurance. The estimation is what it can decide the best to provide Insurance. Customers are having Driving License more those who are not interested in Insurance as compared to customers who are interested in Insurance are least. The data of customer is estimating to get insights very precisely to know each feature role.

Customers with Vehicle age 1-2 years are more likely to interested as compared to the others. Every year it has shown some estimation. When we observe for Vehicle age greater than 2 years have very less chance of claiming Insurance but for less than a year is also less to get Insurance. Annual Premium feature is where customer are very least to pay 500k and above. But looking towards the correlation of each independent variables relation between Age and Policy_Sales_Channel, can observe negative correlation with -0.58. As in most of the feature we have is medium range of correlation (i.e not too positive and not too negative).

According to region code most of region they don't prefer but some of the region code they have slightly high too. Response customer who are eligible for Insurance, Age is varying from 21 - 65. Policy Sales Channel doesn't affect at all at Insurance claiming. Vintage has balanced with (Response : 1 : Customer is interested, 0 : Customer is not interested) but quite less for those Customer who are interested.

As in Age YoungAge - 25 to 45, MiddleAge - 45 to 65 and OldAge - More than 65. We can see that Customers belonging to YoungAge group are more likely not interested in taking the vehicle Insurance but as compared to other Age Group more are interested. Similarly, Region_C in Region Code has the highest chances of not taking Insurance but in Region_A most of them prefer as compared to other who wants Insurance. Channel_A Customers has the highest chances of not taking the Insurance but Channel_C customers wants to claim Insurance as compared to others.

YoungAge Customer prefer more as compared to OldAge and MiddleAge. MiddleAge are also more comparetively more to get likely for Insurance. The customers of YoungAge and OldAge are equally likely to have/not have vehicle Insurance whereas customers of MiddleAge has the highest chances of not having a previously insured vehicle Insurance. The variation is very precise that as the Age increase the Annual Premium also increases as per year. Looking towards vehicle damage 97.9% the vehicle is damage with Response appropriate value of 45,728. While 2.1% we can see the customers vehicle is not damage with Response value of 982.

5. Steps Involved

- **Exploratory Data Analysis**

After loading the dataset I performed this method by comparing our target variable (i.e Response) with other independent variables. This process helps us figuring out various aspects and relationships among the dependent and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the dependent variable.

- **Null values Treatment**

Our dataset doesn't contain any null values, (if it would be which might tend to disturb accuracy and drop them at the beginning of data preprocessing).

- **Extracting features from column**

The Age, Region Code and Policy Sales Channel column was an important feature to gets smaller insights from its.

Extracting YoungAge, MiddleAge and OldAge from Age to create a new feature in dataset. Similarly, getting Region A, Region B and Region C from Region Code. For Policy Sales Channel I've got Channel A, Channel B, Channel C and Channel D. The extraction has made very easy to get useful insights.

- **Univariate and Bivariate Analysis**

The numeric feature has helped with histogram analysis and got insight from heatmap for multi-collinearity. Each feature like Gender, Age, Driving License, Region Code, Previously Insured, Vehicle Age, Vehicle Damage, Annual Premium, Policy Sales Channel etc. shows greater impact on analysis and training the model that how it varies.

The dependent variable has shown relation with every numeric feature of a dataset. The ROC curve and confusion matrix has help to understand each variable as per their respective model.

- **Fitting different models**

  For modeling I tried algorithms like

  1. **Logistic Regression (Implemented using hyper parameter tuning i.e GridSearchcv)**
  2. **XGBoost Classifier (Implemented using hyper parameter tuning i.e GridSearchcv)**

- **Hyper parameter Tuning (GridSearchcv)**

While doing hyper parameters it is important to shrink the coefficient values to get better test accuracy, roc_auc score, f1 score, precision and recall.

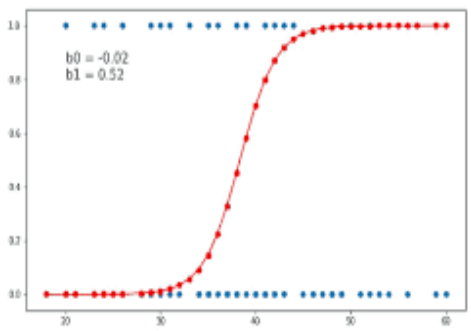
- **Model Evaluation Metrics**

By model evaluation metrics, the calculation value of roc_auc score is important for classification. I've also look for precision, recall, f1 score, and test accuracy to have clear vision of performance of each model. Among all others Logistic Regression was optimal model with roc_auc score of 98%, which is better.

6. Algorithms

1. Logistic Regression:

Logistic regression is a classification technique borrowed by machine learning from the field of statistics. Logistic



Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The intention behind using logistic regression is to find the best fitting model to describe the relationship between the dependent and the independent variable.

Types of Logistic Regression:-

1. Binomial : Where the target variable can have only two possible types e.g:- Health Insurance Cross Sell Prediciton.

2. Multinomial : Where the target variable have three or more possible types, which may not have any quantitative significance. Eg:- Predicting Disease.

3. Ordinal : Where the target variables have ordered categories. Eg:- Web Series ratings from 1 to 5.

For logistic regression, the cost function is

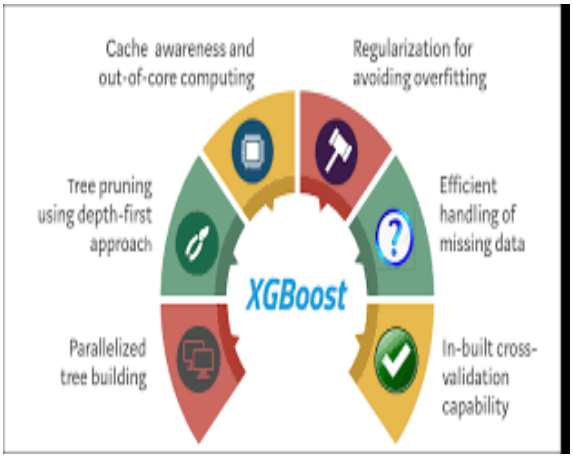
$$\text{Cost}(h_\theta(x), Y(\text{actual})) = -\log(h_\theta(x)) \text{ if } y=1$$
$$-\log(1-h_\theta(x)) \text{ if } y=0$$

given by the equation:

This negative function is because when we train, we need to maximize the probability by minimizing loss function. Decreasing the cost will increase the maximum likelihood, assuming that samples are drawn from an identically independent distribution.

2. XGBoost Classifier :

XGboost, which stands for Extreme Gradient Boosting is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression,



classification and ranking problems.

We've ensemble learning –

- Bagging
- Boosting

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. The boosting ensemble technique consists of three simple steps:

- An initial model F0 is defined to predict the target variable y. This model will be associated with a residual (y – F0)
- A new model h1 is fit to the residuals from the previous step
- Now, F0 and h1 are combined to

$$F_1(x) <- F_0(x) + h_1(x)$$

give F1, the boosted version of F0. The mean squared error from F1 will be lower than that from F0:

$$F_2(x) <- F_1(x) + h_2(x)$$

To improve the performance of F1, we could model after the residuals of F1 and create a new model F2:

$$F_m(x) <- F_{m-1}(x) + h_m(x)$$

This can be done for *'m'* iterations, until residuals have been minimized as much as possible:

Important Parameters which control overfitting and speed:

- n-estimators - the number of runs XGBoost will try to learn.
- learning rate – the shrinkage you do at every step you are making
- max-depth - Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. 0 indicates no limit on depth.

3. Model Performance and

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Model Evaluation Metrics:

- **roc_auc Score**

Compute Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores.

$$ROC - AUC = \int_0^1 TPR(FPR)\, dFPR$$

$$= \int_0^1 TPR(FPR^{-1}(x))\, dx$$

- **Recall**

Recall is a valid choice of evaluation metric when we want to capture as many positives as possible. For example: If we are building a system to predict if a person has cancer or not, we want to capture the disease even if we are not very sure.

```
Recall = (TP)/(TP+FN)
```

- **f1 score**

The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. The F1 score sort of maintains a balance between the precision and recall for your classifier. If your precision is low, the F1 is low and if the recall is low again your F1 score is low.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

- **Accuracy**

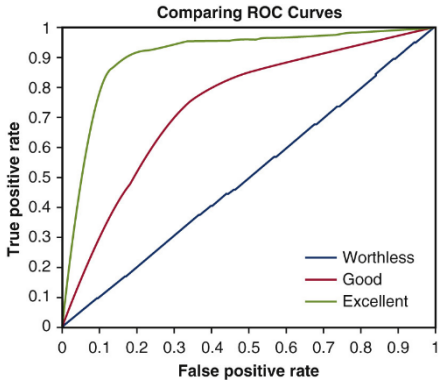
Accuracy is the quintessential classification metric. It is pretty easy to understand. And easily suited for binary as well as a multiclass classification problem.

```
Accuracy =
(TP+TN)/(TP+FP+FN+TN)
```

Accuracy is the proportion of true results among the total number of cases examined.

- **ROC Curve**

ROC curve is almost independent of the response rate. The receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as



its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

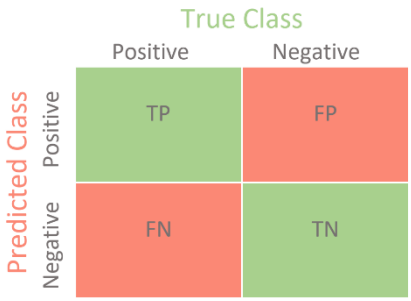
- **Precision**

Precision is a valid choice of evaluation metric when we want to be very sure of our prediction. For example: If we are building a system to predict if we should decrease the credit limit on a particular account, we want to be very sure about our prediction or it may result in customer dissatisfaction.

```
Precision = (TP)/(TP+FP)
```

- **Confusion Matrix**

A confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.



7. Conclusion

That's it! I reached the end of exercise. Starting with loading the data so far I have done EDA, null values treatment, extracting features, some univariate and bivariate analysis, model building and model evaluation metrics. The roc_auc score varies in both the model but optimal model we can find is in Logistic Regression. So roc_auc score of the best model is 98% for this dataset. This performance could be due to various reasons : no proper pattern of data, different model give different accuracy score.

References:-

1. towardsdatascience
2. kdnuggets
3. analyticsvidhya