Insurance Cross-Selling

Ankit Sharma,

Pankaj Kumar,

Soniya kumawat

Data science trainees,

Alma Better, Bangalore

Abstract:

Abstract:

Understanding basic terms and problem domain or space is essential before starting a project. In the financial services industry, cross-selling is a popular term. Cross-selling involves selling free or product as a complement of other product to existing customers. It is one of the highly effective techniques in the marketing. It is a widely used technique to grow business as well as relationship with the customer, also it is much easier to gain profits as getting a new customer is always a little costlier.

1.Problem Statement

In this project, we see that insurance company provides health insurance to its customers. Now, they want to build a model that would help them know if the existing customers who are having health insurance would be interested in buying vehicle insurance or not which would provide by the company.

2. Introduction

An insurance policy is provided by insurer which could be a private or government firm in which an agreement is signed between insurer and insured person where insurer promises to pay guaranteed compensation for incurring loss or damage or illness, death in return of a fix and pre decided premium. A premium is paid by insurer on a regular basis that could be monthly, quarterly, half yearly or annually. Here in this project, we will be talking about a health insurance company who is expanding from health to vehicle insurance as well and want to know if their existing customers would be willing to buy their vehicle insurance policy or not.

For this we will be taking data already provided by Health insurance company of their existing customers.

Data understanding

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. Policy Sales Channel: Anonymized Code for the channel of outreaching to the customer i.e., 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

Data visualization

Exploratory Data Analysis-After loading the dataset we performed this method by comparing our target variable that is Response with other variables. This process helped us figuring out various aspects and relationships among the target and the variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

Conclusion of EDA

- Male members are more interested in insurance plan
- Vehicles with age 1-2 years are interested in taking insurance
- Customers who are interested in Vehicle Insurance almost all have driving license
- persons with vehicle damage only are interested in insurance
- The Feature Vintage has very less information and is Uniformly Distributed, With no skew. Also, the Values are uniformly mixed, in both the classes of the target variable response.
- The individuals with region code 28 are the most as compared to the other ones are interested in insurance
- The most used sales channels are 152, 26 and 124. The best channel that results in customer interest is 152.
- Age and annual premium are highly corelated i.e., 68%
- From the distribution plot we can infer that the annual premium variable is right skewed

DATA preprocessing

Null values Checking-Our dataset does not contain any null values this makes work easy

Encoding of categorical columns -We used label encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features cannot be understood by the machine and needs to be converted to numerical format.
Label encoding is applied on vehicle age, gender, vehicle damage

Balancing unbalance data- As from the distribution of target variables in the EDA section, we know it is an imbalance problem. The imbalance datasets could have their own challenge. For example, a disease prediction model may have an accuracy of

99% but it is of no use if it cannot classify a patient successfully. So, to handle such a problem, we can resample the data. In the following code, we will be using We have used smote to balance the data

Synthetic Minority Oversampling Technique (SMOTE) algorithm working can be explained in 4 steps which are mentioned below: -

- Firstly, we need to choose a Minority class from the sample and then we need to choose a random example from the minority class.
- Then we need to find the **k** of its nearest neighbor, where **k** neighbor is known as argument in **SMOTE** function.
- We choose a random neighbor and afterwards we create a synthetic example at a random point which will be lying in between of the two random points and on the same line joining those points selected.
- The result so obtained would be a convex combination of the two random points selected and these steps may need to be repeated until the dataset is balanced.

• Fitting different models

Data modelling requires different classification algorithms and we tried a few algorithms in our project as well. Following were the classification algorithms we tried: -

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Decision Trees**
4. **KNN**

Switching hyperparameters for improved result and accuracy
It is vital to turn the Hyperparameters of respective algorithm in order to get improved and accurate results and avoid overfitting specifically in the case of Logistic regression.

Algorithms:

1. Logistic regression: -

This is a tool basically used for modelling of statistical data which in its basic form uses a logistic function to solve a binary variable problem.

Its formation is quite like that of linear regression that is why it was given name of ‘regression’.

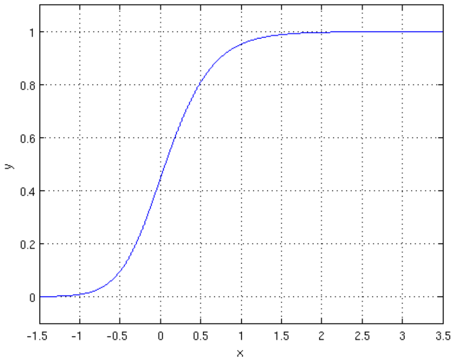
Its application increases when the target variable is categorical in nature.

Functions used in Logistic regression is sigmoid function

Mathematical expression of logistic regression is given by: -

$$f(x)= 1/1+e \,\hat{}\,(-x)$$

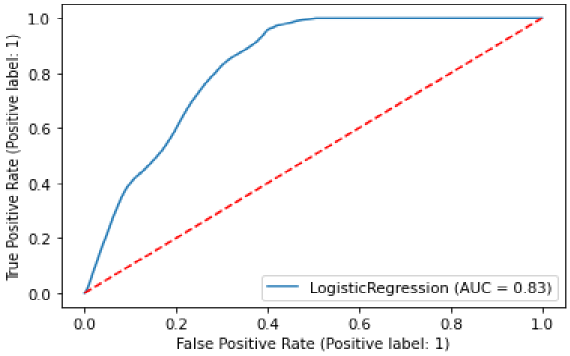
- When we applied random forest, its parameters are showing as follows: -
  Accuracy-88%
  Precision-90%
  Recall-87%
  F1_score-88%



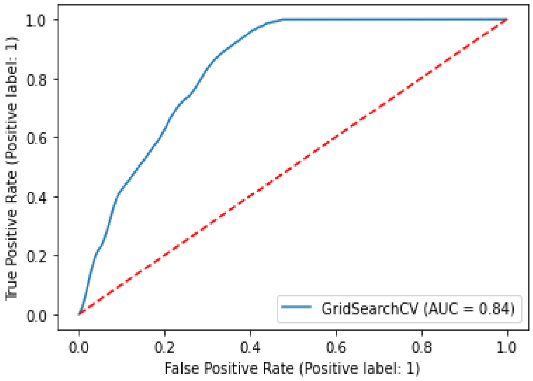
Algorithm used to optimize is **Maximum Log Likelihood**

In logistic regression operation Log likelihood is used in most of the parts, and it is given by: -

$$\ln L(\mathbf{y},\beta) = \ln\prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n}\ln(1-\pi_i)$$

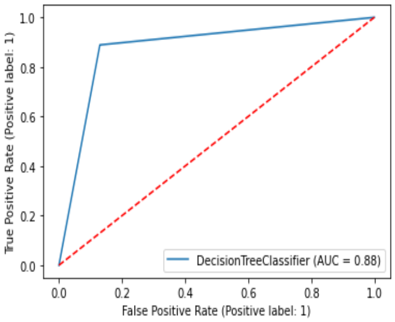


After hyperparameter tuning



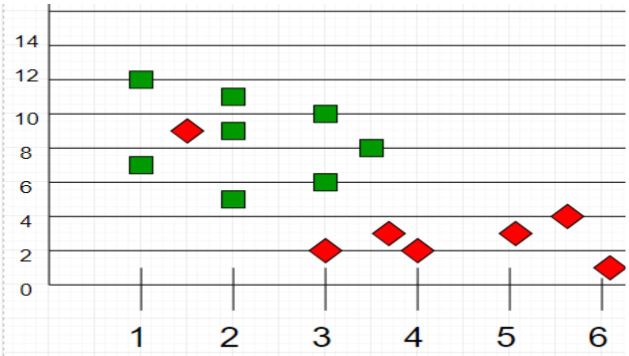
Decision Trees (DTs)

- They are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.
- When we applied Decision tree its parameters are showing as follows: -
- Accuracy- 87%
- Precision-88%
- Recall-87%
- F1_score-88%
- AUC/ROC curve

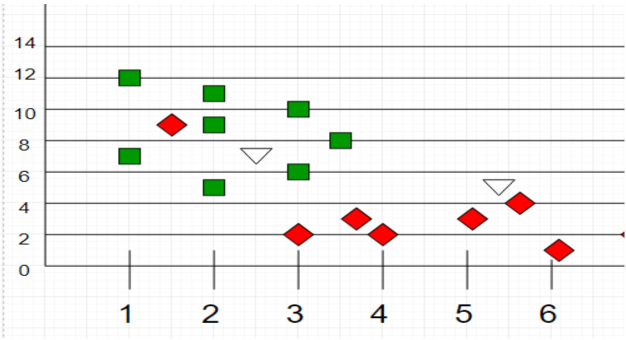


K-Nearest Neighbors

- It belongs to the supervised learning domain We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute. As an example, consider the following table of data points containing two features:

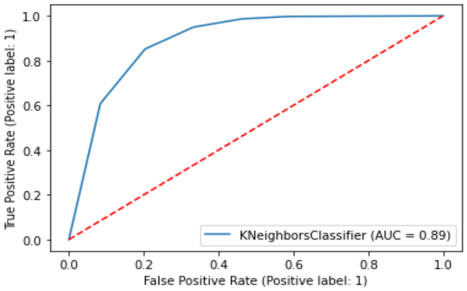


- Now, given another set of data points (also called testing data), allocate these points a group by analyzing the training set. Note that the unclassified points are marked as 'White'.



If we plot these points on a graph, we may be able to locate some clusters or groups. Now, given an unclassified point, we can assign it to a group by observing what group its nearest neighbors belong to.

- This means a point close to a cluster of points classified as 'Red' has a higher probability of getting classified as 'Red'. Intuitively, we can see that the first point should be classified as 'Green' and the second point should be classified as 'Red'.
- When we applied Decision tree its parameters are showing as follows: -
- Accuracy-80%
- Precision-94%
- Recall-74%
- F1_score-83%
- AUC/ROC curve



Basic parameters to evaluate the model performance

1. **Precision**

   Precision is the ratio of correct positive predictions to the overall number of positive predictions: **TP/TP+FP**

2. **Recall-**

   Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: **TP/FN+TP**

3. **Accuracy-**

   Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by: **TP+TN/TP+TN+FP+FN**

4. **F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

   F1 Score = 2*(Recall * Precision) / (Recall + Precision)

5. **Area under ROC Curve (AUC)-**

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance. The higher the area of a curve the higher the accuracy of a model.

Hyper parameter tuning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability, and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV on logistic regression but not much difference seen in performance

1. **Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

Conclusion of Model

After completing this project, we have summarized few facts from the project of insurance cross sell prediction data and the inferences has been confirmed with both the bivariate analysis of each feature,

- The most interested age group in buying vehicle insurance was fund to be 30–60-year age group
- Customers having frequent vehicle damage or vehicle already damaged were keen on buying vehicle insurance
- Customers who owe driving license were much interested in buying vehicle insurance
- Chances of buying insurance highly depends upon the variable given in the project like -Age, Annual premium, Vehicle damage condition, gender.
- After comparing ROC curve, we found out that Random Forest model gives better result than other curves. It gave good accuracy which means the result was very accurate and performance was better than other curves.

References:

1.Analytics vinadya

2.Geeks for Geeks