Technical Documentation

Health Insurance Cross-Sell Prediction

Completed by:

Ankush Kumar

Data Science Trainees @ AlmaBetter, Bangalore

Abstract:

A company agrees to provide a guarantee of compensation for a specific loss, damage, illness, or death in exchange for the payment of a specific premium under the terms of an insurance policy.

In order to attract customers for any insurance policy, a variety of factors are important.

The demographic data presented here includes age, gender, region code, vehicle damage, vehicle age, annual premium, and policy sourcing channel.

We can learn more about the factors influencing news's popularity on social media and get the best classification model by using data analysis and prediction using machine learning models based on the prior trend.

Problem Statement and Objective:

Determining whether a person with an insurance policy would be interested in purchasing car insurance as well. The company can then plan its communication strategy to reach out to those customers and optimise its business model and revenue by building a model to predict whether a customer would be interested in Vehicle Insurance.

This project's goal is to use machine learning algorithms like Logistic Regression, KNN, and Decision Tree to build a predictive model from the provided data set using statistically significant variables.

Introduction

An insurance policy is a contract whereby a business agrees to guarantee compensation in the event of a specific loss, damage, illness, or death in exchange for the payment of a predetermined premium.

The amount of money that the customer must consistently pay to an insurance company in exchange for this guarantee is known as a premium.

Technical Documentation

Vehicle insurance works similarly to medical insurance in that customers must pay an annual premium to the insurance provider company in order for them to be compensated (referred to as "sum assured") in the event that their vehicle is responsible for an unfortunate accident.

Data information:

Data is a crucial component of any effective machine learning model. No matter how good your machine learning models are, without enough rich data, you cannot get a dependable high-performance model from the prediction model.

We have a dataset which contains information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. related to a person who is interested in vehicle insurance. We have 381109 data points available.

Feature Name	Type	Description
id	(continuous)	Unique identifier for the Customer.
Age	(continuous)	Age of the Customer.
Gender	(dichotomous)	Gender of the Customer
Driving_License	(dichotomous)	0 for customer not having DL, 1 for customer having DL.
Region_Code	(nominal)	Unique code for the region of the customer.
Previously_Insured	(dichotomous)	0 for customer not having vehicle insurance, 1 for customer having vehicle insurance.
Vehicle_Age	(nominal)	Age of the vehicle.
Vehicle_Damage	(dichotomous)	Customer got his/her vehicle damaged in the past. 0: Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	(continuous)	The amount customer needs to pay as premium in the year.
Policy_Sales_Channel	(nominal)	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	(continuous)	Number of Days, Customer has been associated with the company.

Response (Dependent Feature)	(dichotomous)	1 for Customer is interested, 0 for Customer is not interested.

METHODOLOGY FOR WHOLE ANALYSIS:

- Exploratory Data Analysis
- Correlation
- Transformation
- Baseline Model
- Performance Metrics
- Optimization
- Hyperparameter Tuning
- Conclusions

Exploratory Data Analysis (EDA):

Data Cleaning and Removal of Duplicity:

- There was not any null value
- And also, no duplicate data was present in our data.

Data Visualization

- We have performed different types of visualization on the basis of Univariate, Bivariate and multivariate analysis.
- Firstly, we have conducted Univariate analysis as we also need to understand various features/columns individually that what kind of importance and insights they bring for our analysis.
- Secondly, we have performed Bivariate analysis so that we can analyse the impact of one column/feature to another feature and where these insights lead us.
- At last, we have performed Multivariate analysis in which we came to know the impact of multiple features.

We have concluded some important and leading insights from our EDA analysis:

Area code 28 has the most health insurance clients and region code 51 has the fewest.
Almost no effect on insurance buy if a person is already having a insurance or not.

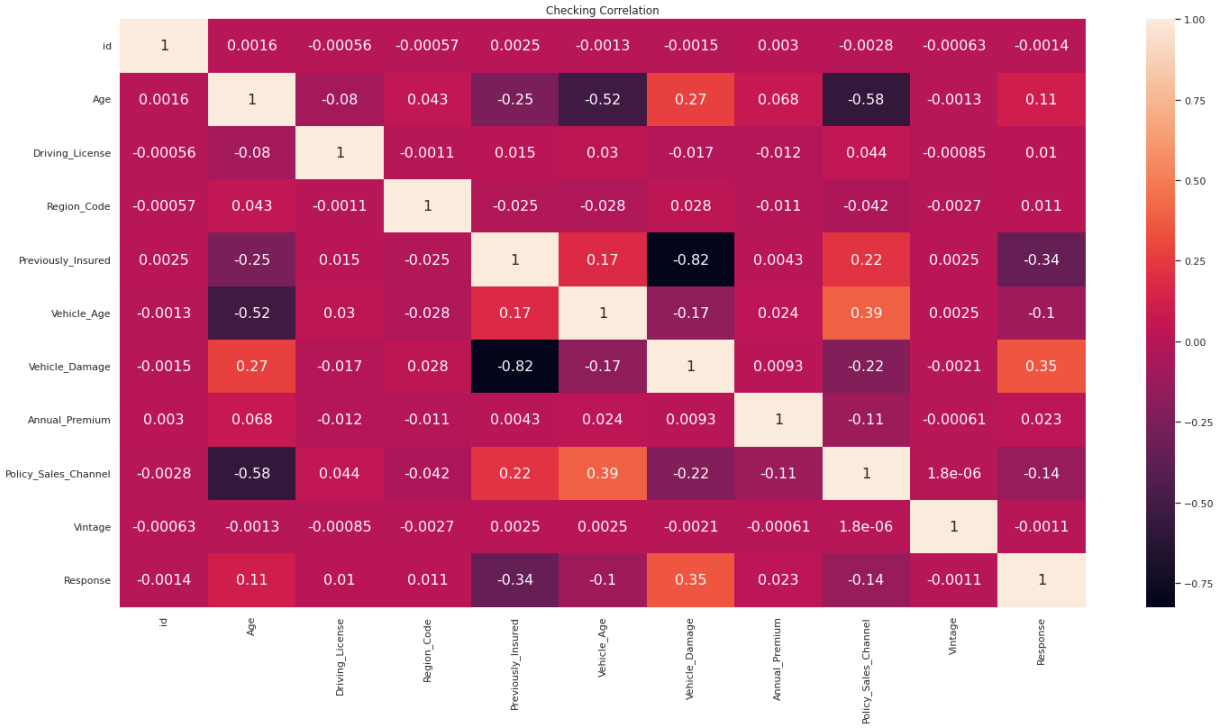
vehicle age between 1-2 years has the greatest number of counts and vehicle age less than 2 years has the least count.

Almost equal count for both the Vehicle property based on damage.

The average age of customers with auto insurance is 34.5 while the average age of customers without auto insurance is 42.4.

Age range of vehicles Customers with newer cars are less likely to be interested in purchasing auto insurance.

Correlation: As we know that correlation is a very important observation task to get the proper idea that how our features are independent from each other and if they are not independent and are related to each other then there will be lots of ambiguity for our model creations that misleads our analysis.



- **<u>As we can see that there is no correlated features</u>**

Feature Transformation:

- We have many features that have some categorical values in them like Gender(Male/Female), Vehicle_Age, Vehicle_Damage

We have applied LabelEncoder from Scikit Learn and transformed these features to numeric format.

Technical Documentation

LabelEncoder:

Sklearn is a very efficient tool for encoding categorical feature levels into numeric values. LabelEncoder encodes labels with values ranging from 0 to n classes-1, where n denotes the number of distinct labels.

Different machine learning models are used to predict the label class and also used Over-sampling technique to balance over imbalanced class label:

Random Forests

Definition:

- The random forest classification technique is made up of several decision trees. When building each individual tree, it employs bagging and feature randomness in an attempt to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree.

Algorithm

- Each tree in the random forest generates a class prediction, and the class with the most votes becomes the prediction of our model.
- A large number of relatively uncorrelated models (trees) acting as a committee will outperform any of the individual constituent models.
- The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't all err in the same direction all of the time).

Decision Trees

Definition:

Decision Tree is a predictive modelling tool that finds ways to divide a data set based on various conditions. It is built using an algorithmic approach. It is among the most popular and useful techniques for supervised learning. A non-parametric supervised learning technique called decision trees is used for both classification and regression tasks. The objective is to learn straightforward decision rules inferred from the data features in order to build a model that predicts the value of a target variable.

Working

- In order to generate the best decision trees possible, the variables that act as nodes in the decision trees are determined by calculating the Information Gain of each independent variable in the dataset.

- To define Information Gain precisely, we must first define entropy, a measure commonly used in information theory that measures the level of impurity in a set of examples.

KNN Classifier:

Definition:

- "K-Nearest Neighbor" is referred to as KNN. It is an algorithm for supervised machine learning. Problem statements involving classification and regression can both be solved using the algorithm. The symbol "K" stands for the number of closest neighbors to a new unknown variable that needs to be predicted or classified.

Working

- Calculate the Euclidean distance between the K neighbors you choose, then choose the K neighbors who are closest to you based on that distance.
- Count the number of data points in each category among these k neighbors, and then assign the new data points to the category where the number of neighbors is highest.

Logistic Regression:

Definition:

- The "Supervised machine learning" algorithm of logistic regression can be used to model the likelihood of a particular class or event. It is applied when the outcome is binary or dichotomous and the data can be linearly separated. That means that binary classification issues are typically addressed by logistic regression.

Working

- Based on one or more predictor variables, logistic regression is used to predict the class (or category) of individuals (x). It is used to simulate a binary result, or a variable with only two possible values, such as 0 or 1, yes or no, or diseased or not.

XGBOOST Classifier:

Definition:

- Extreme Gradient Boosting (XGBoost) is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning library. The top machine learning library for regression, classification, and ranking issues, it offers parallel tree boosting.

Technical Documentation

Working

- It is predicated on the hunch that when previous models are combined with the best possible next model, the overall prediction error is minimized. Setting the desired results for this subsequent model in order to reduce error is the key concept.

SMOTE (used to balance our imbalanced class in our dataset)

Definition:

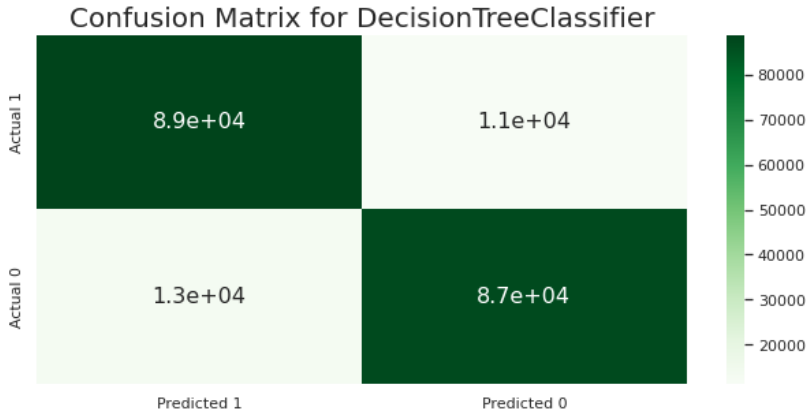
- A statistical method for evenly increasing the number of cases in your dataset is Synthetic Minority Oversampling Technique (SMOTE). The component creates new instances from minority cases that you supply as input that already exist.

Working

- SMOTE finds the k closest minority class neighbors of the minority class instance it has chosen at random, instance a. Next, a line segment in the feature space is formed by joining a and b to form the synthetic instance by randomly selecting one of the k nearest neighbors, b. The two selected instances, a and b, are convexly combined to create the synthetic instances.
- As many synthetic examples of the minority class as needed can be produced using this procedure. It suggests using SMOTE to oversample the minority class in order to balance the class distribution, followed by using random under-sampling to reduce the number of examples in the majority class.

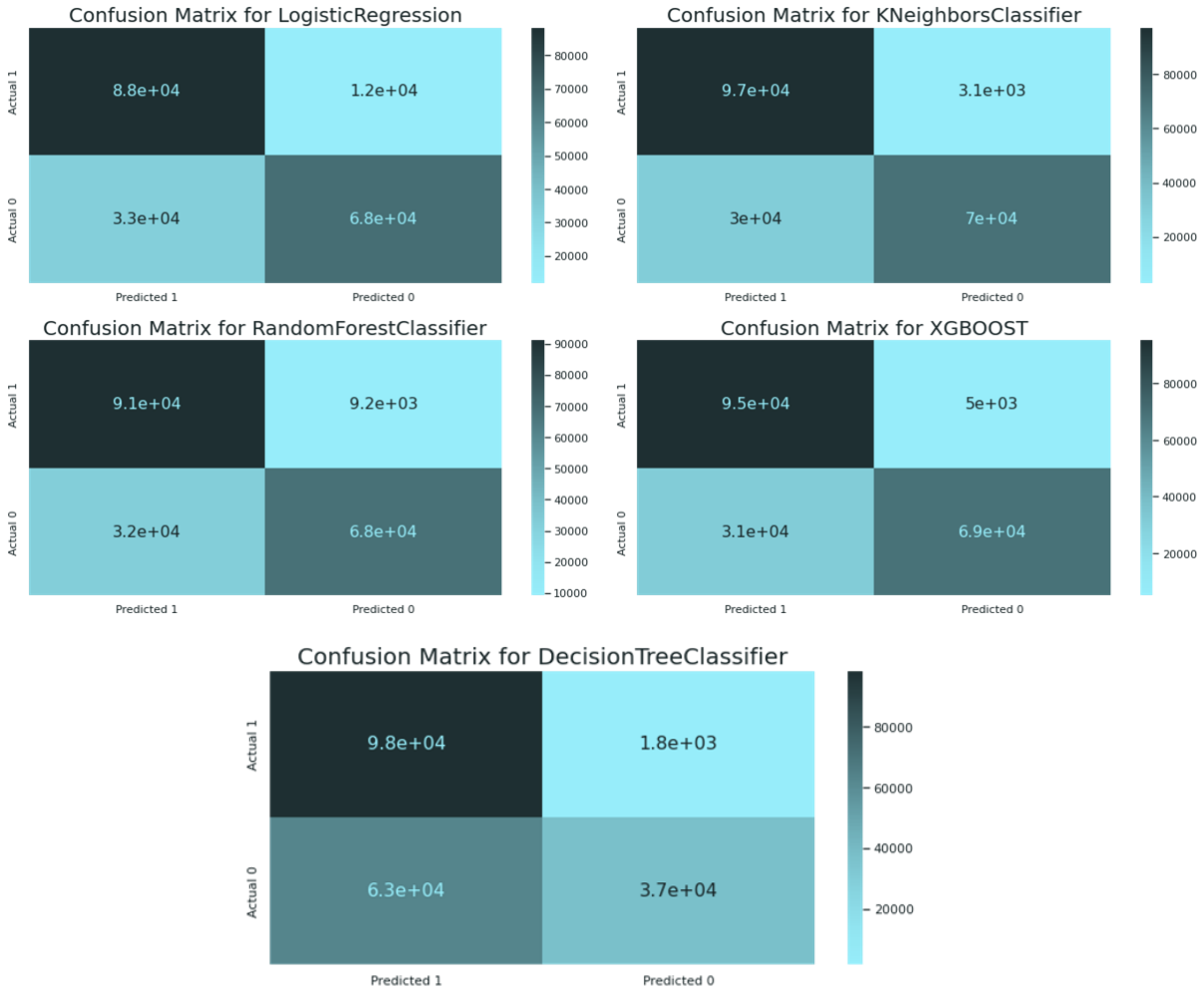
Results:

Observation by applying baseline model:



- False Negative rate of Logistic Regression is 41480, and for Decision Tree it is 11428, which is also very high, and for XGBOOST it is 10978, which is also very high, and we are not considering these models.
- It is 5633 for KNN, which is much lower than these models, and 7468 for Random Forest, which is also lower but higher than the KNN model.
- KNN will be our best model for this entire problem statement.

Optimized Hyper tuned model results:



- KNN's recall rate is now slightly higher than before, but Decision Tree, after hyper parameter tuning, is doing an excellent job and gives us a recall score of 98 percent. ** And yet, logistic regression is still underperforming.
- After hyperparameter tuning, the recall rate of Random Forest drops from 93 percent to 90 percent.
- When we compare KNN and XGBOOST classifiers, we can see that KNN still superior to XGBOOST.

Conclusion and Summary:

- We can see that after hyper parameter tuning, the False Negative rate of Logistic Regression is 11970, which is the highest among all models, and the False Positive rate of Random Forest is 9184, which is also high.
- When we compare the false positive rates of KNN and XGBOOST, we get 3080 and 5001 respectively, which is lower than the Logistic and Random Forest models but higher than the Decision Tree Decision Tree Classifier, which has the lowest False Positive Rate of 1818 of all the models.
- However, if we consider good recall, accuracy, and a low false positive rate, KNN does a great job with (96 percent recall, 83 percent accuracy, and a 3080 false positive rate) if all parameters are considered, our final model should be KNN Classifier.
- However, if we are only concerned with Recall Rate, we should use Decision Tree Regressor.
- As we improve the KNN Classifier Model with more hyper parameter tuning, we will get more desired results that will be more useful.
- We can conclude from our preliminary analysis of feature distribution, outlier detection, and overall, EDA analysis that: Customers who have never had a vehicle damaged are only 0.5 percent interested in vehicle insurance.
- Vehicles less than a year old are more likely to have insurance, with 66 percent insured, while vehicles older than a year but less than two years old are insured in 33 percent.

Future work recommendation:

- Insurance companies should promote and offer more customer-friendly insurance policies.
- Clients with vehicles older than two years are eligible for benefits.
- And more parameters for analysis should be added to make the analysis more explainable.

References:

- Towardsdatascience
- Analyticsvidya
- Becominghumanai
- Siteminder.com
- Tmstudies.net