**Health Insurance Cross Sell Prediction**

**Team members  
Balaji B. Jadhav  
Anant M. Patil   
Nishigandha Ingale**

**Abstract**

This paper predicted a model that indicates whether to buy a car based on primary health insurance customer data. Currently, automobiles are being used to land transportation and living, and the scope of use and equipment is expanding. This rapid increase in automobiles has caused automobile insurance to emerge as an essential business target for insurance companies. Therefore, if the car insurance sales are predicted and sold using the information of existing health insurance customers, it can generate continuous profits in the insurance company's operating performance. Therefore, this paper aims to analyse existing customer characteristics and implement a predictive model to activate advertisements for customers interested in such auto insurance. The goal of this study is to maximize the profits of insurance companies by devising communication strategies that can optimize business models and profits for customers.

**Key words**

Health insurance, cross sell prediction, Automobile insurance.

**1.Introduction**

Today, automobiles have become a key tool for land transportation, and their use range is expanding as a means of living. Cars are a running weapon and have a problem in that accident caused by banks also increase, leading to significant increase in life and property damage. Due to the rapid growth in vehicles, automobile insurance has emerged as a critical business target for insurance companies. It is crucial to expand new contracts by creating new customers to improve business performance. However, as relationship marketing becomes more critical, the repurchase of insurance and maintenance of insurance contracts by existing customers is an important success factor for the insurance industry. In the long term, minimizing the conversion of existing customers to other insurance companies is a positive factor that increases corporate profits. Thus, it became a major problem for insurers to increase business efficiency by maintaining existing contractors by controlling the factors of car purchase conversion.

**Problem Description**

Our client is an 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.

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.

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.

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 optimise its business model and revenue.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

**2. Data Description**

Digging into data we understand that

There are total 11 Features such as

* 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 have DL.
* Region code - Unique code for the region of the customer.
* Previously Insured – 1: customer has already insurance, 0: customer does not have insurance.
* Vehicle age – Age of the vehicle.
* Vehicle damage – 1: customer got his/ her car damage in the past 0: customer didn’t get his/ her car damage in the past
* Annual premium – The amount customer needs to pay as a premium in the year.
* Policy sales channel – Anonymized code for the channel of outreaching the customer. i.e. Different agents, over mail etc.
* Vintage: Number of Days, Customer has been associated with the company
* Response variable is:
* Response:  1: Customer is interested, 0: Customer is not interested

**3.EDA on given Data set**

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

**3.1 Data Analysis:**

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

**3.2 Data Sourcing**

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

**3.3 Data Pre-processing**

A dataset may contain noise, missing values, and inconsistent data; thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

**3.4 Data Cleaning**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

Missing Values

1. Incorrect Format
2. Incorrect Headers
3. Anomalies/Outliers

**3.5 Data Deduplication**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

**3.6 Missing Values**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of service and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.
2. MAR (Missing at random): These values may be dependent on some other features.
3. MNAR (Missing not at random): These missing values have some reason for why they are missing.

**3.7 Dropping Missing Values**

One of the ways to handle missing values is to simply remove them from our dataset. We have known that we can use the is null() and not null() functions from the pandas library to determine null values

* 1. **Handling Outliers**

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter these outliers. The main reason for this detection and filtering of outliers is that the presence of such outliers can cause serious issues in statistical analysis.

There are two types of outliers:

**3.9 Univariate Outliers**

Univariate outliers are the data points whose values lie beyond the range of expected values based on one variable.

* 1. **Multivariate Outliers:**

While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value.

**3.11 Measures of Central Tendency**

The measure of central tendency tends to describe the average or mean value of datasets that is supposed to provide an optimal summarization of the entire set of measurements. This value is a number that is in some way central to the set. The most common measures for analysing the distribution frequency of data are the mean, median, and mode.

**3.12 Measures of Dispersion**

The second type of descriptive statistics is the measure of dispersion, also known as a measure of variability. If we are analysing the dataset closely, sometimes, the mean/average might not be the best representation of the data because it will vary when there are large variations between the data. In such a case, a measure of dispersion will represent the variability in a dataset much more accurately.

Multiple techniques provide the measures of dispersion in our dataset. Some commonly used methods are standard deviation (or variance), the minimum and maximum values of the variables, range, kurtosis, and skewness.

* 1. **Standardizing Values:**

To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale. For example, if the data contains the values of the top speed of different companies’ cars, then the whole column should be either in meters/sec scale or miles/sec scale.

**3.14 Univariate analysis**

Univariate analysis is the simplest form of analysing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

**3.15 Bivariate Analysis**

If we analyse data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

**a) Numeric-Numeric Analysis:**

Analysing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyse it in three different ways.

* Scatter Plot
* Pair Plot
* Correlation Matrix

**b) Numeric - Categorical Analysis:**

Analysing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyse those mainly using mean, median, and box plots.

**3.16 Multivariate Analysis**

Multivariate analysis is the analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behaviour more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

**3.17 Correlation Among Variables**

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour

**3.18 Graphical Representation of The Results**

This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. This is also an essential step as the result analysed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA. Most of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

**4.Algorithms**

**4.1. Decision Tree:**

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called*recursive* partitioning. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.

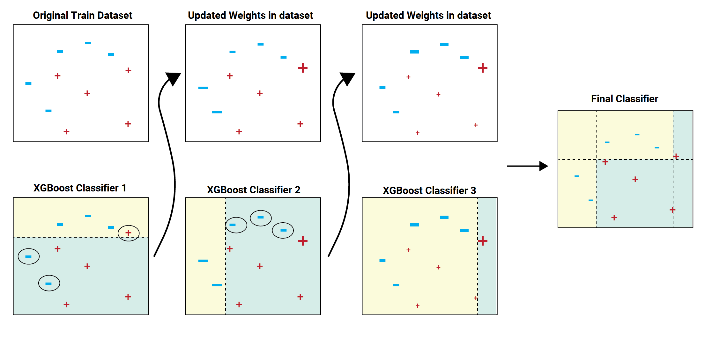


**4.2. Random Forest:**

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the greatest number of times a label has been predicted out of all.

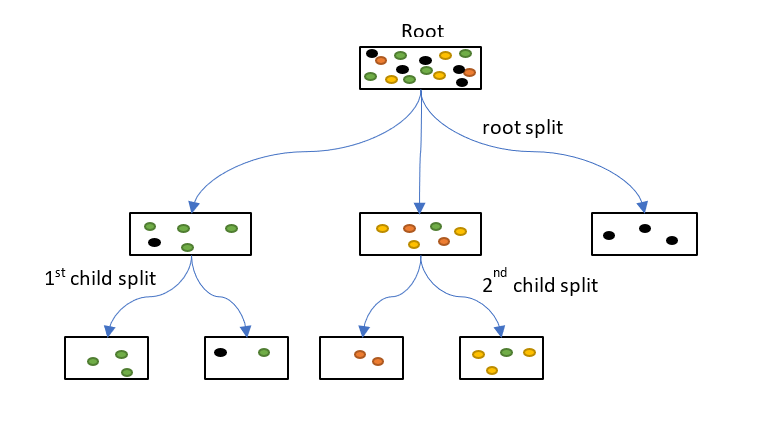
**4.3. XG Boost**

**XG Boost** is an optimized distributed gradient boosting library designed to be highly **efficient, flexible**and**portable**. It implements machine learning algorithms under the [Gradient Boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. XG Boost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.



**4.4. Cat Boost**

Cat Boost is an algorithm for gradient boosting on decision trees. It is developed by Yandex researchers and engineers, and is used for search, recommendation systems, personal assistant, self-driving cars, weather prediction and many other tasks at Yandex and in other companies, including CERN, Cloudflare, Careem taxi. It is in open-source and can be used by anyone.

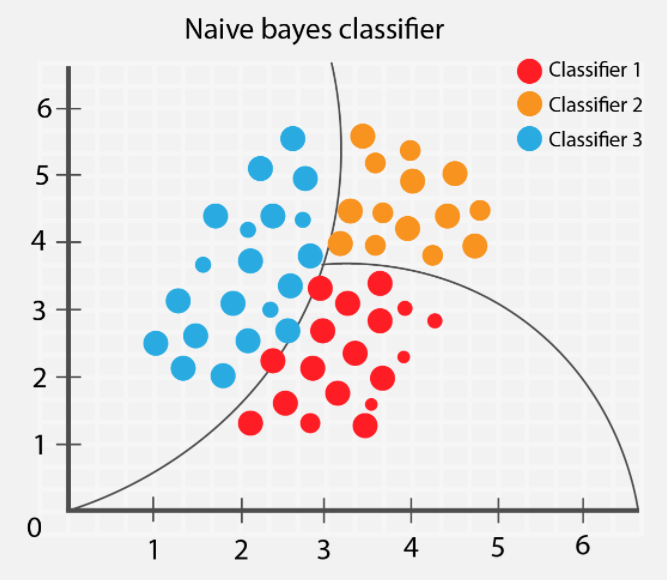


**4.5. Navie Bayes:**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

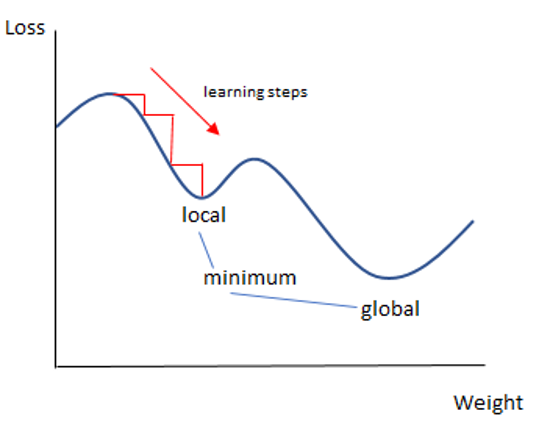
For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.



**4.6. SGD Classifier:**

SGD Classifier is a Linear classifiers with SGD training. This estimator implements regularized linear models with stochastic gradient descent (SGD) learning: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule. SGD allows minibatch (online/out-of-core) learning, see the partial fit method. For best results using the default learning rate schedule, the data should have zero mean and unit variance. This implementation works with data represented as dense or sparse arrays of floating point values for the features. The model it fits can be controlled with the loss parameter; by default, it fits a linear support vector machine (SVM)



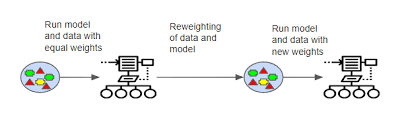
**4.7.KNN Classifier:**

K Nearest Neighbour (KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won’t vote. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.



**4.8. Gradient boosting classifiers:**

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets. The Python machine learning library, Scikit-Learn, supports different implementations of gradient boosting classifiers, including XGBoost. In this article we'll go over the theory behind gradient boosting models/classifiers, and look at two different ways of carrying out classification with gradient boosting classifiers in Scikit-Learn.



**5.Conclusions**

After implementing various models on the given data such as Logistic Regression, Decision Tree Classifier, XG boost, Naïve Bays Classifier, SGD Classifier, Cat Boost Classifier, KNN Classifier, Gradient Boosting Classifier, AdaBoost Classifier, Random Forest Classifier. We get maximum accuracy with Decision Tree Classifier, Random Forest Classifier, XG boost but in case of Decision Tree and Random forest accuracy is decreased for testing data it is the case of Over fitting. In case of XG Boost accuracy decreases but with less percentage here there is no problem of overfitting. XG boost with train accuracy as 0.88 and test accuracy of 0.81

**6.References**

[1] “A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree” by Su Hyun AN1 , Seong Hee YEO2 , Minsoo KANG3

[2]”A study on the meaning of automobile in the no insurance automobile injury insurance” by Choi, B. G

[3]”Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction” by Wagner A. Kamakura