HEALTH INSURANCE CROSS SELL PREDICTION

Sunanda Debnath, Ajay Tiwari,

Data science trainees,

AlmaBetter, Bangalore

Abstract: -

All of this analysis predicted a model that suggests cross-selling is a great way to make more money for any insurance company without starting from scratch. Whether you are buying a car based on primary health insurance customer data. We can build a business from the book you already have with your current customer relationships. It is beneficial not only for the company but also for the clients. It's not just about more profit. But it's about integrating value and bringing solutions to the indemnity issues your customers face. Just like health insurance, there is car insurance, where every year the customer has to pay a certain amount of premium to the insurance company, so that in the event of an unfortunate vehicle accident, the insurance company provides compensation (so-called 'certain amount') to the customer. Our dataset is based on a database of health insurance customers. This experiment can help to understand what can influence the cross-selling factors of insurance plans for existing customers. The model can be used for any insurance plan data set to predict cross-selling.

***Keywords:* machine learning, Supervised machine learning, Cross-selling, Predictive Model building, Logistic Regression, Random Forest , XGB classifier**

1. Problem Statement

Today, cars have grow to be a key device for land transportation, and their use as a method of residing is converting extra and extra. Cars are a not unusual place weapon and feature a break in that injuries as a result of banks are at the upward thrust together, ensuing in a huge boom in harm to lifestyles and property. Due to the fast increase of vehicles, coverage has grow to be a key enterprise goal of coverage organizations. It is important to enlarge new contracts with the aid of using obtaining new clients to boom enterprise performance. As dating promoting will become very essential, the acquisition of coverage and the preservation of rules with the aid of using present clients is a essential achievement element withinside the coverage industry. In the lengthy run, minimizing the conversion of present clients to different coverage organizations may be a advantageous trouble that will increase the earnings of organizations.

Coverage is an arrangement with the aid of using which an organization undertakes to offer a warranty of emolument for a chosen loss, damage, illness, or death in reciprocation for the price of a chosen top class. A top class is a sum of profit that the client needs to pay usually to an insurance organization for this warranty.

The dataset consists of these important features: -

1. **ID**: Unique ID for the customer
2. **Gender**: Gender of the customer
3. **Age**: Age of the customer
4. **Driving License**: whether the Customer has DL
5. **Region Code**: Unique code for the region of the customer
6. **Previously Insured**: Whether Customer already has Vehicle Insurance
7. **Vehicle Age**: Age of the Vehicle
8. **Vehicle Damage**: Whether a customer

got his/her vehicle damaged in the past.

9. **Annual Premium**: The amount customer pays yearly for Insurance.

10. **Policy Sales Channel**: Anonymized Code for the channel of outreaching out 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

2. Introduction

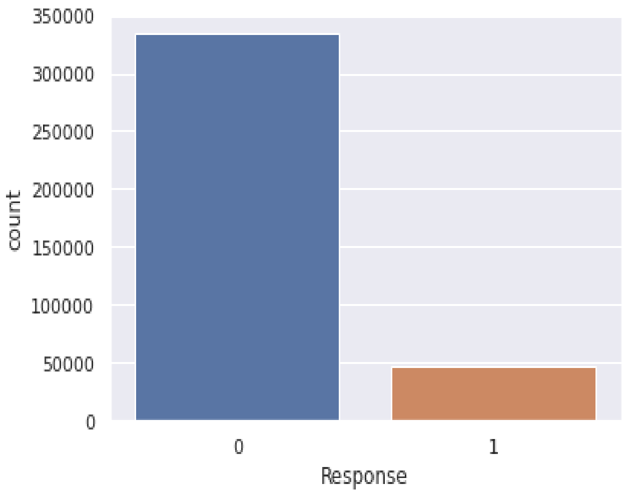
Cross-Selling is a brand new advertising method primarily based totally on information analysis, which located that exceptional desires exist as well, who can end up clients and meet their desires via income of diverse associated offerings or products.

We are utilizing the Logistic Regression algorithm, RandomForest algorithm and XGBClassifier for building different prediction models, RandomForest algorithm build nodes based on certain decision and hence can be very useful for random data that has no specific distribution. Logistic Regression is a parametric algorithm and hence certain properties are to be verified for good results. The Random Forest engenders decision trees on randomly selected data samples, gets predictions from each tree, and selects the best solution by means of voting. Here we have features like age, driving license, and vehicle age which is going to give a relationship with the response variable.

3. Response type:

1. Interested
2. Not Interested

Segregation of clients in step with their interests is the first step closer to the construction of the prediction model, we have features like age, vehicle age, and the annual top rate of the coverage taken. These functions have a correlation with the reaction variable.



As you can see on the graph, there are very few interested customers whose stats are less than 50000 and which customers are not interested those is above 300000

4. Need for cross-selling Model Building:

The need for Cross-Selling Models is:-

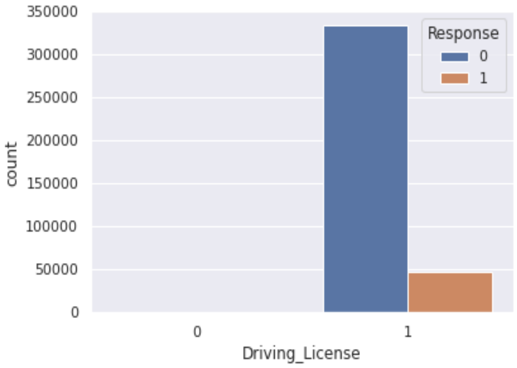
1. For higher to recognize the clients, wants.
2. To pressure income maximization from already present clients as opposed to launching new advertising and marketing campaigns
3. To offer a non-public experience to the clients.
4. To make an approach focused on the patron phase to maximize earnings.
5. To make the adjustments in the goods if required for a higher market.

5. Steps involved:

Exploratory Data Analysis

Exploratory data analysis (EDA) is utilized by statistics scientists to investigate and inspect statistics units and summarize their major characteristics, frequently using statistics visualization methods. It enables decide how exceptional to govern statistics reassets to get the solutions you need, making it less complicated for statistics scientists to find out patterns, spot anomalies, take a look at a hypothesis, or take a look at an assumption

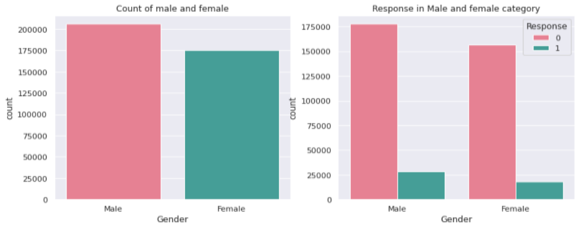
Customers who are interested in Vehicle Insurance almost all have driving license.



The number of male is greater than 200000 and The number of female is close to 175000.

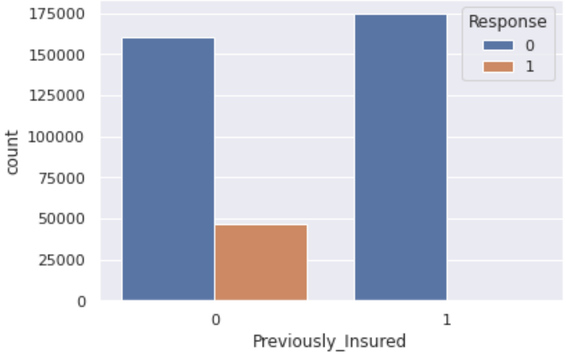
• The number of male is interested which is greater than 25000 and The number of female is interested which is below 25000.

• Male category is slightly greater than that of female and chances of buying the insurance is also little high

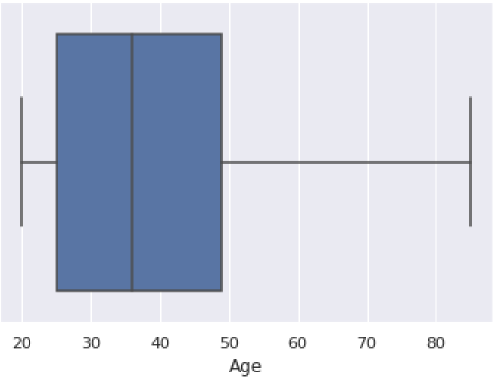


• Customer who are not previously insured are likely to be interested.

• 1 indicate no one interested and the value of this is null.



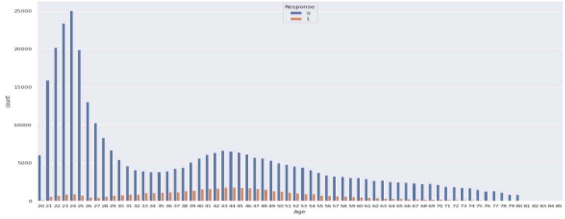
also we cheak for outliers.



Young people below 30 are not interested in vehicle insurance. Reasons could be lack of experience, less maturity level and they don't have expensive vehicles yet.

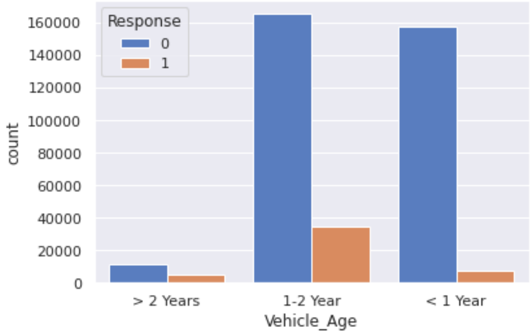
• People aged between 30-60 are more likely to be interested.

• From the boxplot we can see that there no outlier in the data.

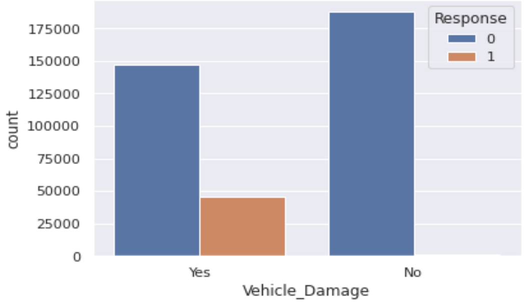


• Customers with vehicle age 1- 2 years are more likely to interested as compared to the other two.

• Customers with Vehicle Age <1 years have very less chance of buying Insurance.

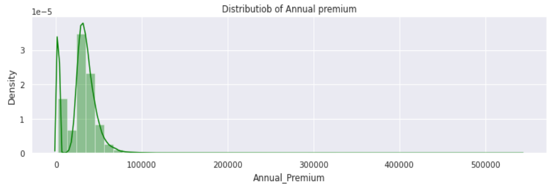


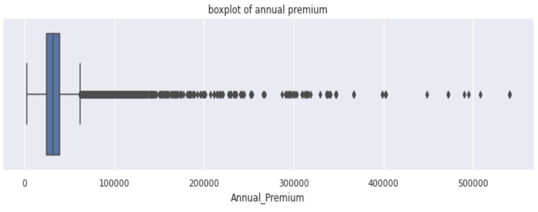
we can infer that if the vehicle has been damaged previously then the customer will be more interested in buying the insurance as they know the cost.



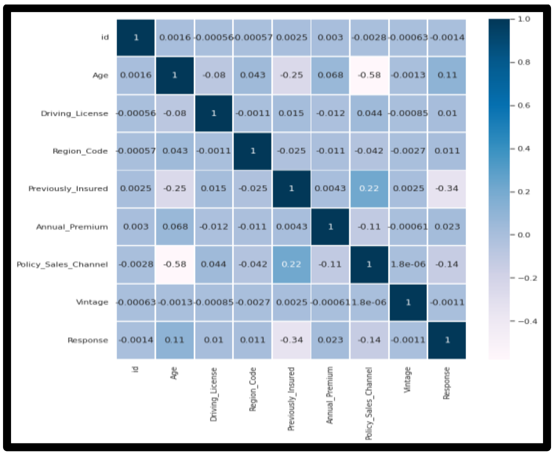
- From distribution plot we can see annual premium is rightly skewed.

- From boxplot we can see there are too much outliers.





As we can see the heat map graph Vehicle Damage column is more correlated with target variable.



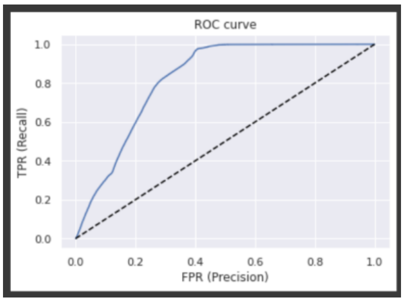
Feature Engineering:

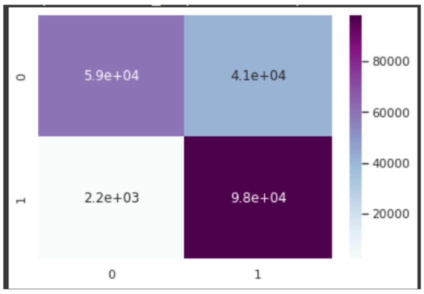
Feature engineering is the method of selecting, manipulating, and reworking uncooked statistics into functions that may be utilized in supervised gaining knowledge. For the system to gain knowledge of carrying out nicely on new tasks, higher functions might also additionally want to be designed and trained. As you could know, a "feature" is any measurable enter that may be utilized in a predictive model - it may be the color of an item or the sound of a person's voice. Feature engineering, in reality, put, is the act of changing uncooked observations into preferred functions through the usage of statistical or system-gaining knowledge of approaches.

- **Model building, Predictions, and Forecasting**

1. Logistic Regression:

Logistic regression is a sort of easy system learning that plays regression estimation on proportional, proportional, or specific data. Compared to more superior type and regression techniques, it's miles primarily based totally on a totally easy theory, however, probabilistic evaluation is viable for specific data. In different words, it's miles viable to expect the prevalence of a particular occasion via way of means of the use of an unbiased variable that has an immediate effect on the established variable. Logistic regression evaluation can provide an explanation for the forms of institutions and interactions resulting from the version shape and might evaluate the effect of explanatory variables on reaction values via parameter inference. In addition, due to the fact it's miles viable to carry out discrimination and type primarily based totally on anticipated chance, numerous industries along with medicine, telecommunications, and finance are acting duties to expect the chance of an occasion going on the use of logistic regression evaluation.





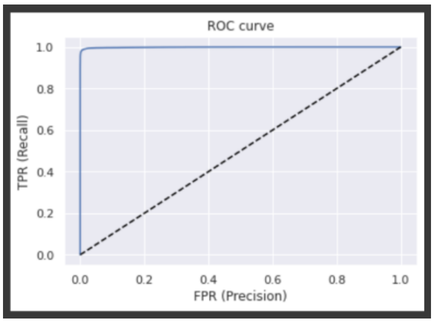
2. Random Forest :

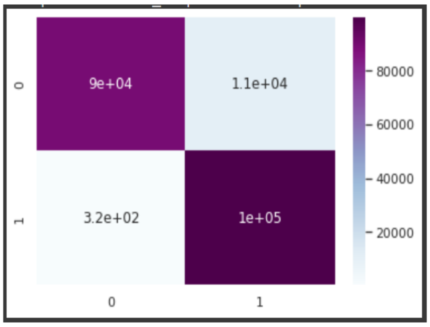
Random Forest is a famous machine learning algorithm that belongs to the supervised learning technique. It may be used for each Classification and Regression issue in ML. It is primarily based totally on the idea of ensemble getting to know, that's a procedure of mixing a couple of classifiers to remedy complicated trouble and enhance the overall performance of the model.

As the call suggests, "Random Forest is a classifier that carries some of the choice bushes on diverse subsets of the given dataset and takes the common to enhance the predictive accuracy of that dataset."

Instead of counting on one choice tree, the random wooded area takes the prediction from every tree and is primarily based totally on the bulk votes of predictions, it predicts the very last output.

The extra quantity of bushes withinside the wooded area results in better accuracy and forestalls the trouble of overfitting.



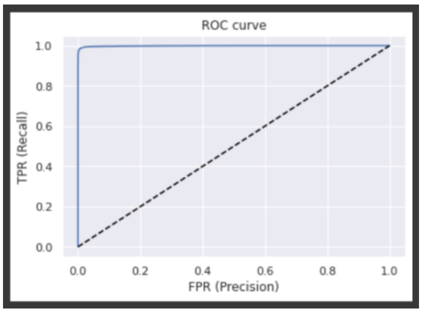


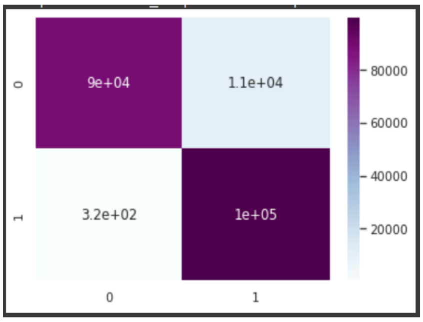
3. XGB classifier :

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.

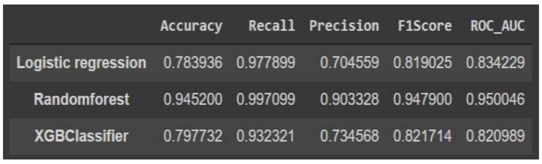
It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset's features.





Combining all regression models:



	Accuracy	Recall	Precision	F1Score	ROC_AUC
Logistic regression	0.783936	0.977899	0.704559	0.819025	0.834229
Randomforest	0.945200	0.997099	0.903328	0.947900	0.950046
XGBClassifier	0.797732	0.932321	0.734568	0.821714	0.820989

6. Challenges :

Identify a highly imbalanced data set and manage it carefully.

Ensure that the model treats all groups fairly.Set prediction thresholds before deploying the model. After deployment, understand the behavior of the model on real data. It is difficult to perform predictive analysis for end users.

7. Conclusion :

87.7% of clients spoke back as No to purchasing car insurance. It surely indicates that a maximum of the clients aren't interested in shopping for car insurance.

• Males are 30% more likely to reply as sure for car coverage than females. So business enterprises may want to attention greater on concentrated on male clients and do greater promotions centered toward the woman clients.

• Most of the clients have riding licenses and out of them, 12% are probable to reply as sure for car coverage.

• There isn't any factor in attaining out to clients who have already got car coverage as nearly they all spoke back negatively about purchasing every other coverage.

• 22% of clients who spoke back

definitely do not have preceding coverage. So, the business enterprise ought to attention greater such clients as conversion opportunities are better in such cases.

• Company ought to attention to clients whose car is greater than 2 years old, as 30% of instances they may be interested in shopping for coverage, which is large in comparison to different features.

• Customers with cars aged much less than 12 months are least interested in insurance as at the same time as shopping for a car human beings frequently purchase 1 year coverage. Companies should not spend greater time on those clients as simply 4% of instances they may be probable to mention Yes for car insurance.

• Customers who have broken their motors in beyond are greater sensitive toward shopping for car insurance. In truth 24% of instances, they spoke back definitely primarily based totally on this dataset.

7. References

1. Stack Overflow!
2. GeeksforGeeks
3. Analytics Vidhya
4. Almabetter
5. GitHub
6. Towards data scienc

Conclusion:-

- Due to the Response variable's value 1 being much lower than its value 0, the provided dataset is an imbalanced dataset.
- Compared to their female counterparts, male consumers own a little bit more vehicles and have a higher likelihood to get insurance.
- Customers between the ages of 30 and 60 are the most likely to get insurance whereas Vehicle insurance is not interesting to anyone under the age of 30. The lack of involvement, a lack of knowledge about insurance, and possibly the lack of expensive vehicles are potential causes.
- Customers with driving licenses are more likely to purchase insurance
- Compared to consumers with vehicles less than one-year-old, those with vehicles between one and two years old are more interested in purchasing insurance.
- Due to their personal experience with the costs associated with vehicle repairs, customers with vehicle damage are more likely to purchase insurance.
- The variable such as Age, previously insured, and Annual premium is more affect the target variable.
- We used different types of algorithms to train our model like Logistic Regression, Random Forest model, Decision tree, and XGB Classifier. And Also, we tuned the parameters of the XGB Classifier and Random Forest model. Comparing the model on the basis of precision, recall, accuracy, and F1 score we can see that the XGB Classifier model performs better. Even comparing the ROC curve XGB Classifier performed better because curves closer to the top-left corner indicate