HEALTH INSURANCE CROSS SELL PEDICTION

Manoj Patil M
Saksham Tripathi
Gulzar
Bindu Kovvada
Deepak Kumar Gautam
Data science trainees,
Alma Better, Bangalore

Abstract:

- The objective was to anticipate Customers who are interested in purchasing vehicle insurance.
- Exploratory Data Analysis is done on the dataset to get the insights from the information however the principal invalid qualities are taken care of. Likewise, some hypothesis testing was additionally performed from the experiences from EDA.
- After that Response variable is our objective variable must be highlighted where Analysis activities are performed on it and after that visualization has done for it to understand hidden insights.
- From that point forward, all that was left was to track down the important factors and feature encoding and fit our models by creating various features, and further, the model is assessed utilizing the metrics.

1.Problem Statement

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.

For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for upto Rs. 200,000. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only Rs. 5000/-
, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2- 3)
would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

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 un

fortunate accident by the vehicle, the insura nce provider company will provide a compe nsation (called ‘sum assured’) to the custom er.

Building a model to predict whether a custo mer would be interested in Vehicle Insuran ce is extremely helpful for the company bec ause it can then accordingly plan its commu nication strategy to reach out to those custo mers and optimise its business model and re venue.

Now, in order to predict, whether the custo mer would be interested in Vehicle insuranc e, you have information about demographic s (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, s ourcing channel) etc.

2.Attribute Information:

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 h ave DL, 1: Customer already has DL

5. Region_Code: Unique code for the regio n 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/he r vehicle damaged in the past. 0: Customer didn't get his/her vehicle damaged in the pa st.

9. Annual_Premium: The amount customer needs to pay as premium in the year

10. PolicySalesChannel: Anonymized Cod e for the channel of outreaching to the custo mer ie. Different Agents, Over Mail, Over Phone, In Person, etc.

11. Vintage: Number of Days, Customer ha s been associated with the company

12. Response: 1: Customer is interested, 0: Customer is not interested

3. Introduction

- insurance is a contract, represented by a policy, in which an individual or entity receives financial protection or reimbursement against losses from an insurance company.

- This EDA will use Python libraries, matplotlib, and Seaborn to examine the Subscribed health insurance customers dataset through visualizations and graphs.

- The dataset is of Subscribed Health insurance customers from insurance companies, contains information such as Age, Gender, Driving Licence, Region Etc.

    - Machine learning has a wide range of applications in our organization. Prediction and analysis have long been the most well-known application of machine learning, which fuels our sale prediction.

    - We're also utilizing machine learning to assist in designing our Sale strategies and Campaign programmes by identifying traits that lead to successful content.

    - We utilize it to help Companies to rapidly expand their reach to customers with appropriate data driven decisions.

- We can also employ machine learning to improve Service and Customer retention, Target oriented promotions.

- Our major goal in this project is to identify Customers who are interested in purchasing vehicle insurance based on subscribed health insurance data of the company.

4. EDA:

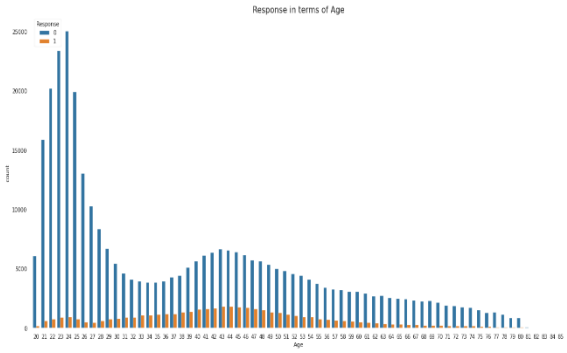
- **Observing and Exploring Dataset**

Exploratory data analysis is a method with help of this we can understand dataset, we can create some insights from data. We can understand statistics part of the data like mean, mode etc.

With eda we can find null values, missing values, duplicate in the data set and outliers. We can find correlation between features in dataset. In eda we can perform various data visualization method on data and observe what is happening in data.

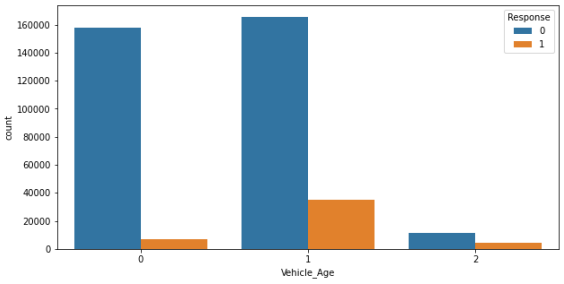
After observing the data, we would say that there are 12 columns and 381109 rows.

Age vs Response



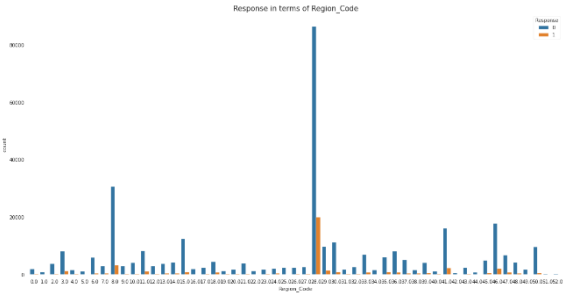
- We can see that Ages below 30 are not more interested in purchasing vehicle insurance may be because lack of experience and maturity levels.
- People who are above 30-55 are more likely to be interested.
- After 50 the lines are declining that's means age above 50 are less interested.

Vehicle age Vs Response



Vehicle age between 1-2 years customers are more interested in insurance than other two.

Region code Vs Response

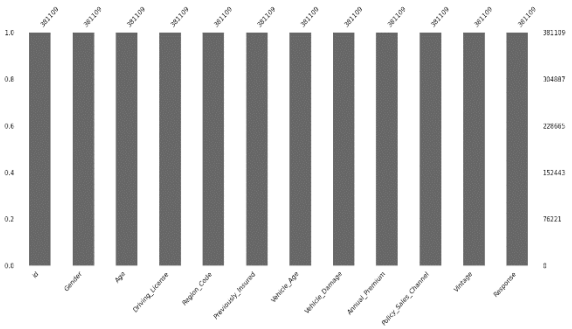


Region code 28 have more customers.

4. Feature Engineering:

- **Null, Missing and Duplicate Values Treatment**

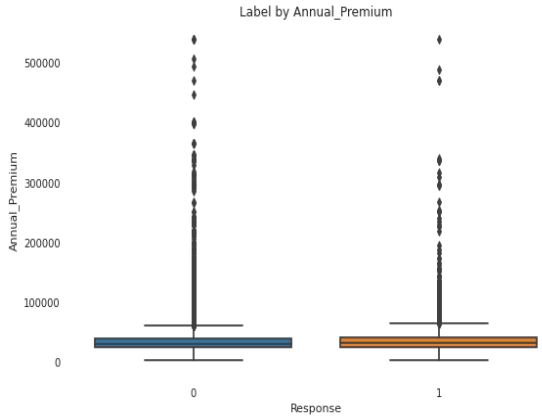
It is an important aspect of Data Cleaning because there can be some null, missing and duplicate values in our dataset. But our dataset doesn't contain a null or missing values which might tend to disturb our accuracy, if it has null or missing values then we have to drop them at the beginning of our project in order to get a better result.



- **Outliers handling**

  Checking outlier in the dataset because Outliers is also something that we should be aware of. Why? Because outliers can markedly affect our models and can be a valuable source of information, providing us insights about specific behaviours. Outliers is a complex subject and it deserves more attention

  We have a lot of outliers in our Premium column, driving license and response column but due to less business knowledge we are not removing any outliers.



- **Encoding of categorical columns**

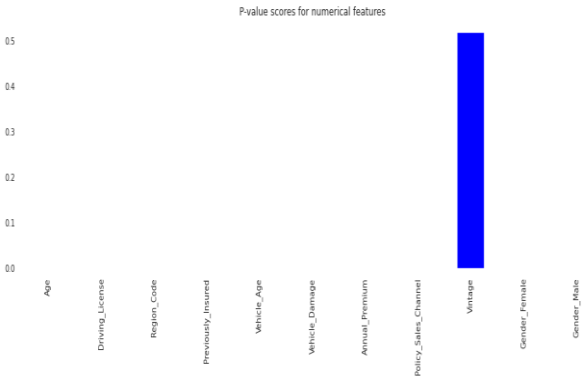
  We used One Hot and label Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

5. Feature Selection:

For feature selection we used three method -

1. **VIF Method-** In this method we didn't get good score so we skip(commented) it.

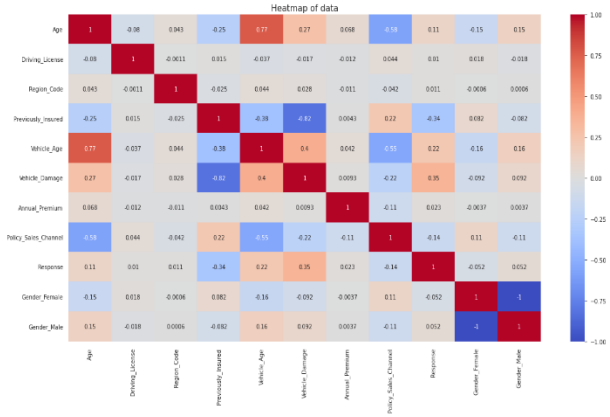
2. **F-classify-**This method gives us promising results so we move ahead with F-classify method and removed some unnecessary feature vintage.



3. **Pearson correlation-**This method gives us correlation between all the features from this method we observe *Gender_female and male 100% Mult icollinearity we can remove any one fe ature among these 2

   * Previously insured, vehicle_age and vechicle_damage have high correlation s with dependent variable

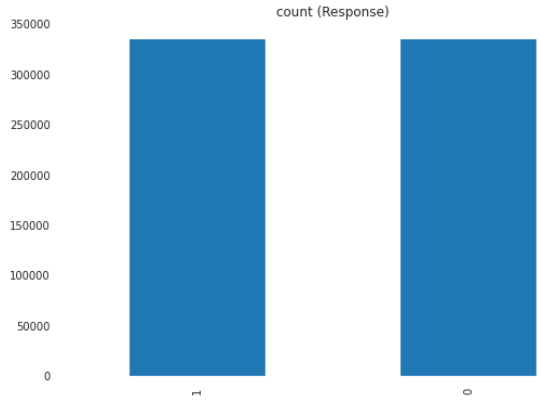
   * Vintage has verly less negative corre lation with dependent variable



6. Preparation for Model Making

- **Imbalance technique-**

  One of the major issues when dealing with unbalanced datasets relates to the metrics us ed to evaluate their model. Using simpler m etrics like accuracy score can be misleading In a dataset with highly unbalanced classes, the classifier will always "predict" the mos t common class without performing any ana lysis of the features and it will have a high a ccuracy rate, obviously not the correct one. We have tried with undersampling, oversa mpling, and SMOTE. Of these, oversampli ng gives the best result.



- **Splitting –**

  train test split is a model validation procedure that allows you to simulate how a model would perform on new/unseen data.
  In this particular step we splitted our data to train and test data with 30% test data.

- **Standardization of features**

  Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.
  The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

We used two scaler Min-max and Standard for our data but Standard scaler gives good results so we proceed further with it.

7. Making Models:

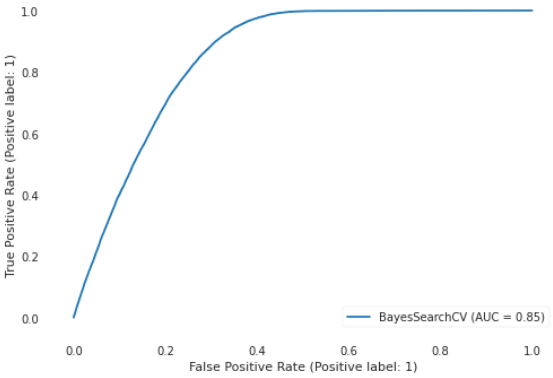
	Name	Train_Time	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
0	LinearClassifier	0.797006	0.757691	0.664734	0.707163	0.253370	0.879645	0.803622	0.784030	0.393026	0.757691	0.758792
1	LogisticRegression	3.525643	0.784026	0.639722	0.765400	0.254498	0.975423	0.977221	0.810722	0.400441	0.784026	0.784178
2	GNB	0.106923	0.784022	0.638713	0.765396	0.254493	0.975423	0.977221	0.810719	0.400035	0.784022	0.784173
3	BNB	0.258527	0.786640	0.645063	0.769572	0.254691	0.973707	0.975802	0.820275	0.403948	0.786640	0.787105
4	KNeighborsClassifier	1.190680	0.903740	0.766629	0.860280	0.286320	0.964054	0.598567	0.909216	0.387353	0.903740	0.694411
5	DecisionTreeClassifier	2.476098	0.980540	0.620970	0.979439	0.294516	0.980031	0.324297	0.980040	0.300898	0.980540	0.607545
6	RandomForestClassifier	76.680098	0.980527	0.625212	0.979364	0.307778	0.979488	0.334729	0.980640	0.320688	0.980527	0.614447
7	GradientBoostingClassifier	55.353739	0.786079	0.701276	0.730096	0.282570	0.924043	0.825135	0.826668	0.432822	0.786079	0.787470
8	XGBRFClassifier	19.065524	0.783766	0.637365	0.794867	0.259102	0.976332	0.978002	0.810683	0.399478	0.783766	0.783865
9	AdaBoostClassifier	16.152070	0.796545	0.668917	0.731119	0.275971	0.930089	0.930610	0.821772	0.426533	0.796545	0.796216
10	LgbmClassifier	7.662085	0.804956	0.701302	0.741012	0.283366	0.937015	0.930954	0.827000	0.434483	0.804956	0.799866

Here we used different types of classification algorithm to compare which one is giving best result and accuracy. It is a sales data so our most important metric is Recall. So with accuracy we also need to focus on algorithm which is giving best recall. As we see boosting algorithms are giving promising results. Also Random Forest Classifier giving good accuracy. Lets go for hyperparameter tuning.

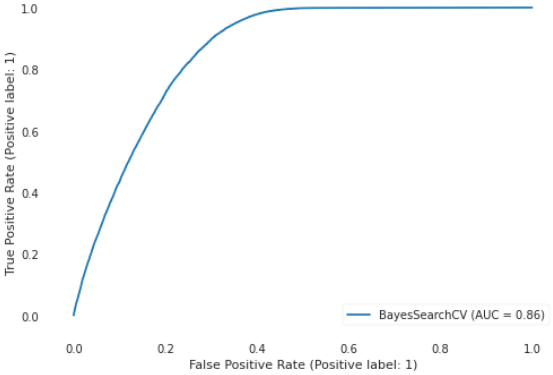
8. Hyperparameter Tuning:

Now we performed hyperparameter tuning on the models which are performing well.

Random Forest Classifier: We used Bayes search Cv for hyperparameter tuning. We used max_depth,min_samples_leaf,min_samples_split,n_estimators and max_features as our hyperparameter to tune our model. After tuning it is giving accuracy around 0.73 and recall around 0.79 which are better.



Lgbm Classifier: : We also used Bayes search Cv for hyperparameter tuning of Lgbm Classifier. We used max_depth,num_leaves,min_split_gain,n_estimators and n_jobs as our hyperparameter to tune our model. After tuning it is giving accuracy around 0.71 and recall around 0.80 which are good.



9. Final Result:

The ML model for the problem statement was created using python with the help of the dataset, and the ML model created with LGBM and Random Forest models performed better than other models.

In comparison to both models, the LGBM model performed well on the most essential evaluation metric, 'Recall,' with values of 0.82 on train data and 0.80 on test data. As a result, **we conclude Lgbm Classifier is the best model for this dataset.**

10.Conclusion:

Our client is an insurance firm that has supplied Health Insurance to its customers. They now need assistance in developing a model to predict whether the policyholders (customers) from the previous year will be interested in the company's Vehicle Insurance.

Building a model to predict if a client is interested in Vehicle Insurance is extremely beneficial to the company because they can then plan communication strategy to reach out to those customers and optimise its business model and revenue.

Now, we have information about demographics (gender, age, region code type), vehicles (vehicle age, damage), policies (premium, sourcing channel), and so on to predict whether the customer would be interested in Vehicle insurance.

Key points:

- Customers aged between 30 to 60 are more likely to buy insurance.
- Youngsters under 30 are not intrigued by vehicle insurance. Reasons could be the absence of involvement, less awareness about insurance and they may not have costly vehicles yet.
- Consumers with 1-2-year-old vehicles are more interested as compared to others.
- Consumers with less than 1-year-old Vehicles have very less chance of buying Insurance
- Customers with Driving License have a higher chance of buying Insurance.
- Customers with Vehicle_Damage are likely to buy insurance.
- The male category is somewhat more noteworthy than that of females and chances of purchasing

the insurance are likewise minimally high.

- The variable such as Age, previously insured, Annual premium is more affecting the target variable.
- Comparing the ROC curve, we can see that the Random Forest model performs better. Because curves closer to the top-left corner indicate better performance.
- We can see that **LGBM model perform better** for this dataset.

Improvements:

1. By using a marketing and advertising approach, we can reduce the gender gap.
2. We can clearly see that we have a larger number of consumers without vehicle insurance, therefore we can easily target them directly with our campaign.
3. Since there are less policy holders with vehicles older than two years, we must pay more attention to the other two categories (1-2 years and >1 year). Because most sales agencies that offer vehicle insurance for the first year are actually our target and we can give them the best incentives to reduce competition in the market.
4. As we saw that we have nearly equal policy holders for both vehicle damage status, so we can target those policy holders whose vehicles are damaged in the past

11.References:

- Stackoverflow
- GeeksforGeeks
- Kaggle
- Machinelearningmastery
- Stack exchange