|  |
| --- |
| Almabetter |
| **HEALTH INSURANCE CROSS SELL PREDICTION** |
| Capstone project - 3 |
|  |
| **Manoj Patil M**  **Saksham Tripathi**  **Gulzar**  **Bindu Kovvada**  **Deepak Kumar Gautam**  **Data science trainees,**  **Alma Better, Bangalore** |
| **7/18/2022** |

|  |
| --- |
| Our client is an insurance company that has provided Health Insurance to its customers now they need our 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. |

**1.Problem Statement**

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.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 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. PolicySalesChannel:  Anonymized Code for the channel of outreaching to the customer ie. 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

**3. Introduction**

As we all knew Cross-selling involves selling customers related items when they are making a purchase. It's important not only because it boosts revenue, but also because it increases customer satisfaction, builds engagement, and helps to create solid and lasting customer relationships.

Cross-selling insurance allows you to earn additional profit without the cost of searching for new leads. Additionally, you build your client relationship by staying up to date on events and changes in your client's lives which might require new or greater coverage. This can lead to improved client retention.

Our client is an Insurance company that has provided Health Insurance to its customers now they need our 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.

After getting your model reports company can do some Digital marketing and automation to grow insurance agency in the field of vehicle insurance through cross-selling and increase their revenue via investing in reproductive marketing campaign.

**4. EDA:**

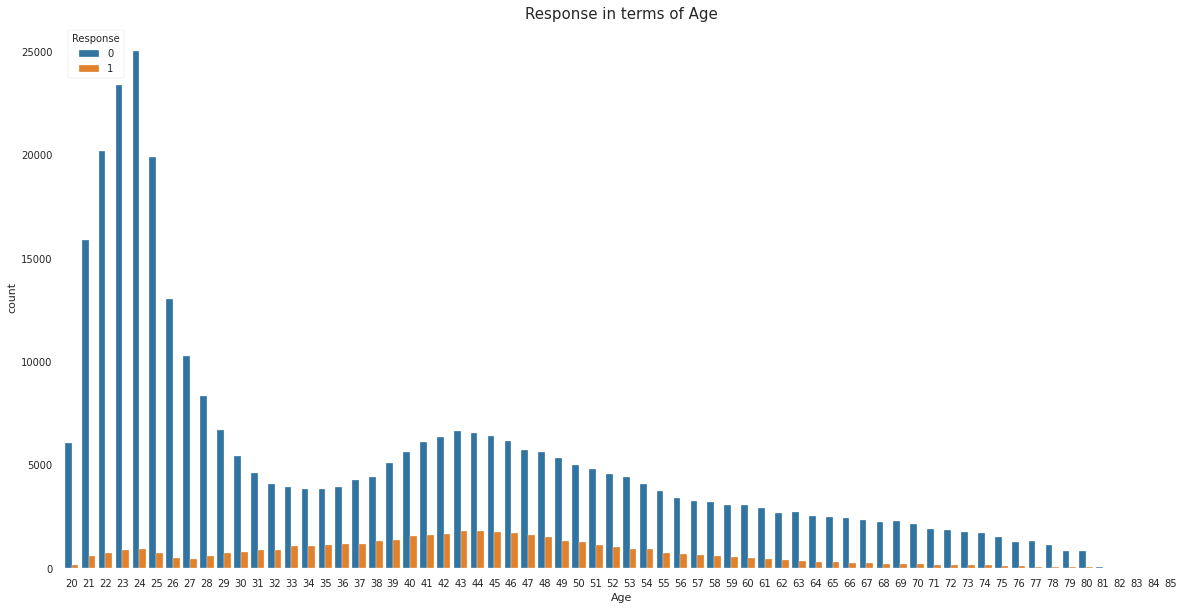
* **Observing and Exploring Dataset**

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.

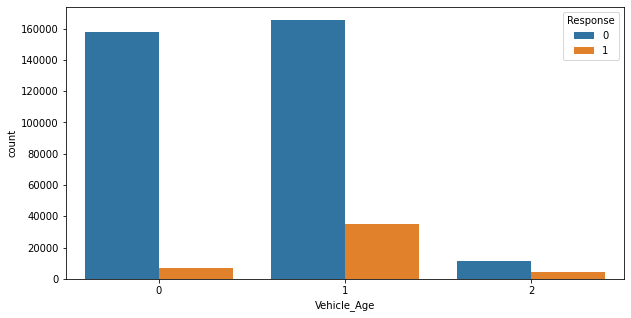
With EDA we can find null values, missing values, duplicate in the data set and outliers. We can find correlation between features in dataset. In EDA we can perform various data visualization method on data and observe what is happening in data.

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

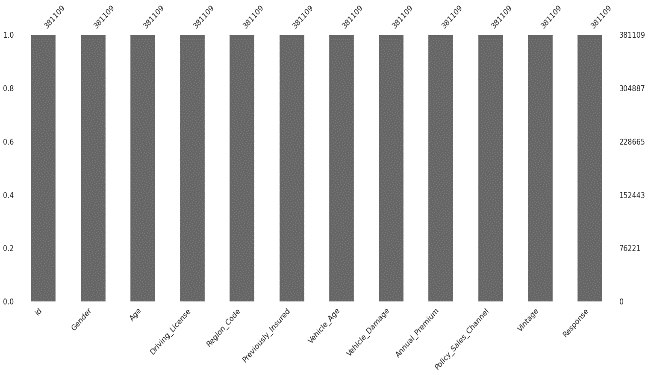
### 

### Region code 28 have more customers.

### 5. Feature Engineering:

### Null, Missing and Duplicate Values Treatment

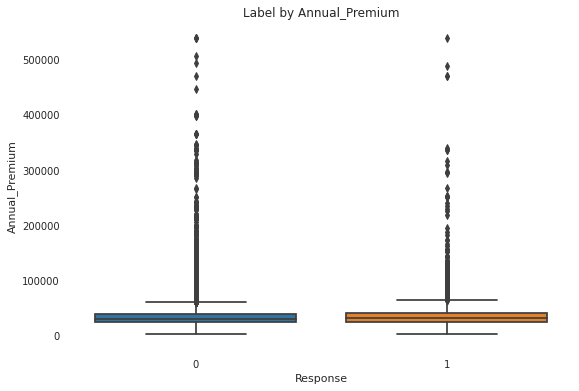
### It is an important aspect of Data Cleaning because there can be some null, missing and duplicate values in our dataset. But our dataset doesn't contain a null or missing values which might tend to disturb our accuracy, if it has null or missing values then we have to drop them at the beginning of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



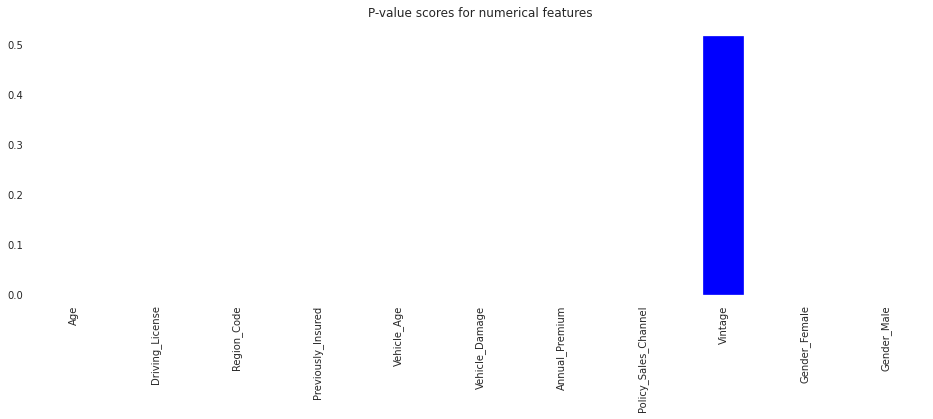
* **Encoding of categorical columns**

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.

**6. Feature Selection:**

For feature selection we used three method -

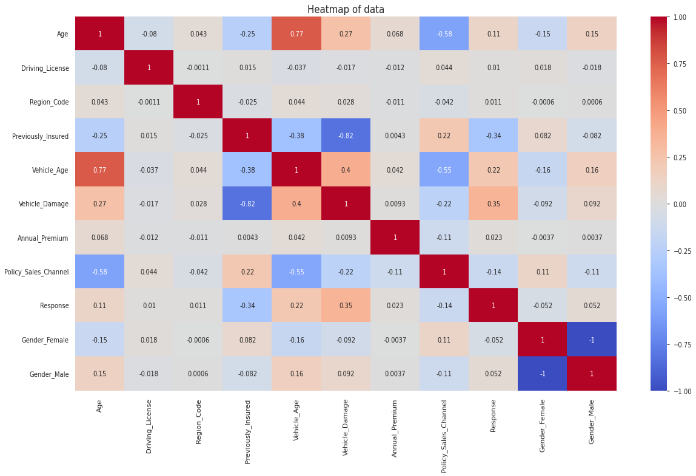
1. **VIF Method-** In this method we didn't get good score so we skip(commented) it.
2. **F-classify-**This method gives us promising results so we move ahead with F-classify method and removed some unnecessary feature vintage.



1. **Pearson correlation-**This method gives us correlation between all the features from this method we observe \*Gender\_female and male 100% Multicollinearity we can remove any one feature among these 2

\*  Previously insured, vehicle\_age and vechicle\_damage have high correlations with dependent variable

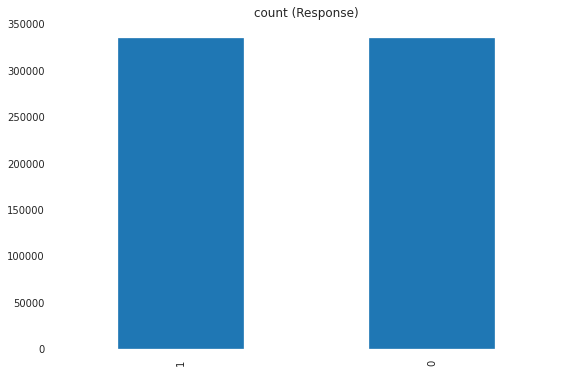
\*  Vintage has verly less negative correlation with dependent variable



**7. Preparation for Model Making**

* **Imbalancetechnique-**One of the major issues when dealing with unbalanced datasets relates to the metrics used to evaluate their model. Using simpler metrics like accuracy score can be misleadingIn a dataset with highly unbalanced classes, the classifier will always “predict” the most common class without performing any analysis of the features and it will have a high accuracy rate, obviously not the correct one.

We have tried with undersampling, oversampling, and SMOTE. Of these, oversampling 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.

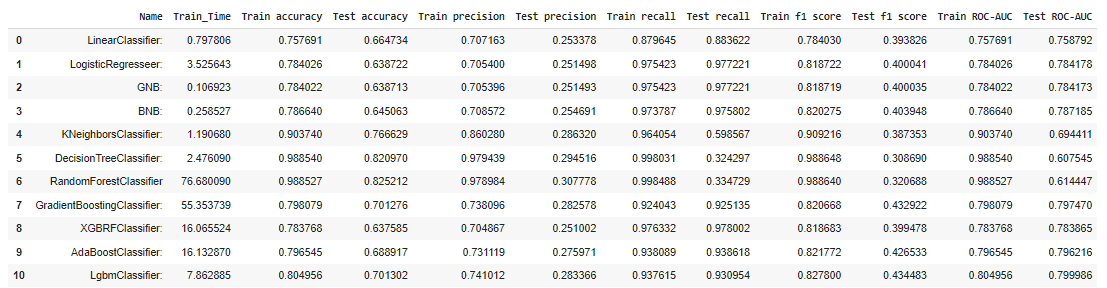
* **Standardization of features**

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.

**8. Making Models:**

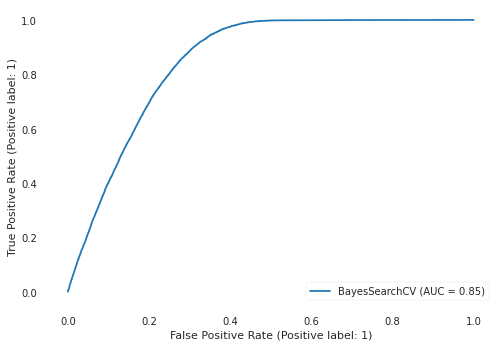
****

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. Let's go for hyper-parameter tuning.

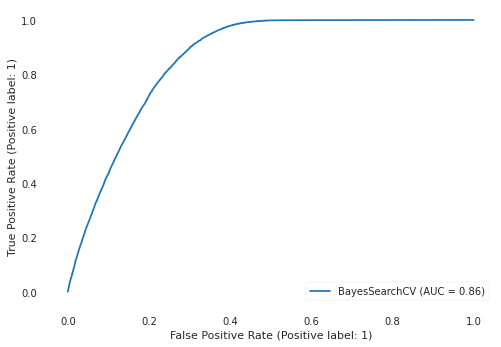
**9. Hyper-parameter Tuning:**

Now we performed hyper-parameter 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.



**10. 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 0.95 on train data and 0.92 on test data. As a result, **we conclude Lgbm Classifier is the best model for this dataset.**

**11. Inferences and Conclusion:**

Starting with loading the data so far we have done EDA , null values treatment, dropping unnecessory columns, outliers handling, visualization, knowing the distribution, feature engineering,Applying some sampling technique(US, OS and SMOTE), model making, finalizing our best model and then we do some hyper-parameter tuning also.

The Lgbm Classifier was the best model when compared with rest all models for this data set. For all the models This Classifier worked the best because it has highest recall in comparison to other models which is important to us in this project..

**It gives 0.95 recall on train and around 0.92 recall on test data for positive response which can be good enough.**

**Key points:**

* We have some gender gap that female is less intrested than male for getting insurance for their veh
* Customers with **Vehicle age between 1 and 2 years** are **more likely** to interested.
* Customer **who are not insured previously** are **more likely** to be interested.
* Age, Previously\_insured, Annual\_premium are having a large predictive power.

**Improvements:**

1. By using a marketing and advertising approach, we can reduce the gender gap.
2. we can easily target those policy holders whose vehicle age is between 1-2 years via target audience of our ads compaign.
3. we can leave those policy-holders out who have already insured thier vehicle or we can focus them for next year target with some discount on price.

**12.References:**

* Stackoverflow
* GeeksforGeeks
* Kaggle
* Machinelearningmastery
* Stackexchange