Contents

[Introduction 3](#_Toc100525930)

[Problem Statement 4](#_Toc100525931)

[Literature 4](#_Toc100525932)

[Data 5](#_Toc100525933)

[**Data Description** 5](#_Toc100525934)

[Data Exploration: 5](#_Toc100525935)

[**Data Transformation:** 6](#_Toc100525936)

[**Heat Map for correlation:** 6](#_Toc100525937)

[Predictive Modeling 7](#_Toc100525938)

[**Prepare for modeling:** 7](#_Toc100525939)

[**Fixing the Imbalanced Dataset:** 7](#_Toc100525940)

[**Data Split:** 8](#_Toc100525941)

[Logistic regression model: 8](#_Toc100525942)

[**Lasso & Ridge Regression -** 8](#_Toc100525943)

[**Evaluation Matrix -** 9](#_Toc100525944)

[Classification Trees - Random Forest: 9](#_Toc100525945)

[Findings 10](#_Toc100525946)

[**Modeling conclusion:** 10](#_Toc100525947)

[Highlights of Data Analysis: 10](#_Toc100525948)

[Correlations: 11](#_Toc100525949)

[Implications/recommendations for business 11](#_Toc100525950)

[References 12](#_Toc100525951)

[Appendix 12](#_Toc100525952)

# **Introduction**

The Health insurance industry generated 15.5 billion USD in economic activity and pays over 209 million USD in state taxes. (CT ERC, 2019).

ABC is an insurance company that has been selling health insurance to its customers. An insurance policy is an agreement between the company and the customer, wherein the company agrees to provide compensation to the customer in case of any loss or damage. The details of the agreement are sealed in a contract that is signed by both parties in exchange for a premium.

In the case of an insurance company, cross-selling insurance is when the company sells other or additional insurance products to existing customers. It shreds down the cost of acquiring new clients in addition to increasing profits. There are several benefits to cross-selling insurance:

1. Reduces the cost and efforts for marketing to a new customer base
2. Increases customer lifetime value by adding another touchpoint in their business relationship with the company
3. Brand growth as an insurance company
4. Bundling products provide greater satisfaction to the consumer and makes their life easier
5. On the off-set, it builds a great advantage in extending leads in the long-term through customer recommendation as a one-stop insurance hub

ABC is looking to cross-sell vehicle insurance to existing policyholders of health insurance from the past year. ABC has data pertaining to various attributes of their existing customers and is looking to predict what category and how many customers would be willing to also purchase vehicle insurance from them. Although several policies of health insurance cover costs of auto accidents, these policies do not cover damage to vehicles alone. There are also certain clauses that require exclusive auto insurance coverage to spill into full damage coverage for any unforeseen car accidents. Understanding which types of customers are likely to buy vehicle insurance will aid ABC in starting conversations early. Cross-selling opportunities are also very time-sensitive. It is important to introduce the product in the early phases of a customer’s lifecycle with the company in order to convert leads into sales. The knowledge of which customers are more likely to buy the cross-sell product will help ABC in making the best use of this time-sensitive attribute of cross-selling insurance policies.

# **Problem Statement**

1. To build a model that predicts whether a Health Insurance customer will purchase an Auto Insurance from ABC Insurance.
2. It would be beneficial for the company to plan its communication strategy appropriately in order to reach out to those customers and optimize its business model and revenue.

# **Literature**

Cross-selling is a popular concept among insurance companies, many make efforts to gain predictive insights from their private datasets. Typically, the insights derived from these datasets are not available to the general public. However, certain businesses report a massively positive impact from using predictive analytics for insurance cross-selling. LexisNexis claims to have increased policy conversions for a large insurance company by 246 percent through predictive analytics.

We got references from Kaggle regarding the analysis performed by users, however, the goal was particularly different for us. The dataset we had included limited columns however, we performed Exploratory Data Analysis in depth so that the prediction model can have better accuracy. We specifically performed transformation to the columns that would give more business value. In the modeling part, we specifically chose Logistic Regression and Random Forests as our models because our focus was aligned with more explainability of the model. The recommendations that we have suggested are totally unique with the problem statement.

We also reached out to a few stakeholders to understand the market needs and this was the challenging part for us as a team. However, we kept searching for the relevant columns that can be used to generate more precise insights from the data.

# **Data**

## **Data Description**

ABC is using the data for its existing health insurance customers for the past year. The training set includes 381109 rows/customers and 12 columns/attributes. The test set includes 127037 and 11 attributes. The description of each variable is as follows:

* **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 already has DL
* **Region\_Code:** Unique code for the region of the customer
* **Previously\_Insured 1 :** Customer already has Vehicle Insurance, **0 :** Customer doesn't have Vehicle Insurance
* **Vehicle\_Age:** Age of the Vehicle
* **Vehicle\_Damage 1 :** Customer got his/her vehicle damaged in the past. **0 :** Customer didn't get his/her vehicle damaged in the past.
* **Annual\_Premium:** The amount customer needs to pay as premium in the year
* **PolicySalesChannel:** Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
* **Vintage:** Number of Days, Customer has been associated with the company
* **Response 1 :** Customer is interested, **0 :** Customer is not interested

# **Data Exploration:**

Upon initial exploration of the dataset, we found that there were no missing values and no null values in the dataset. (**Appendix 1.7**)

We generated count plots from the seaborn library for python for the response variable and found that the dataset was imbalanced as the responses for 0 were 334399 and the responses for 1 were only 46710. **(Appendix 1.8**)

We also generated count plots for other variables. The count plot for region\_code (AR) displays significantly strong participation from Region 28 (**Appendix 1.1**). The count plot for policy\_sales\_channel displays especially important insights as it displays a trend in the response from different sales channels that can approximately be divided into two bins (**Appendix 1.3**).

We can see from the box plot (**Appendix 1.6**) that there is a rise in the value of a particular column, Annual Premium, from 75% to the maximum amount. The presence of outliers may have an impact on the model's performance and accuracy. There are three distinct strategies for overcoming outliers based on column skewness. We have the right skewness in our scenario; hence the LOG transformation was the best option. With the aid of IQR, we were able to normalize the column range (Interquartile Range).

## **Data Transformation:**

Based on the results of the data exploration we transformed our variables to suit the needs of the models to be tested.

Encoding is used to transform data into an appropriate format so that it can be optimally used in the system. In this dataset, we had several variables that needed transformation.

1. The gender variable that pertained to two categories – male and female, was encoded to integers – 1,0.

2. For vehicle\_damage, yes and no were encoded to 1 and 0.

3. Vehicle\_age column that pertained to < 1 year, 1-2 years and >2 years, was encoded to 1,2 and 3.

4. Bins were created for the annual\_premium column : <24,000, 24000 – 40000, and >40000.

5. Bins were created for the age column : <= 27, >=28, and >=36. These were encoded to 1,2,3.

6. Bins were created for the policy\_sales\_channel: 26 – 124 and 152 – 156. These were encoded to 1,2.

7. Dummy variable created for Region 28 from the column region\_code as data exploration hinted at a significantly different relationship of Region 28 with the target variable.

8. Dummy variable created for >2 in vehicle\_age based on the results of data exploration.

## **Heat Map for correlation:**

A heat map provides a graphical source of representation that is color-coded to represent the strengths of correlation between various variables. After transforming our variables and creating new columns for necessary dummy variables, we did a heat map analysis to check for correlation between variables. Correlation can be helpful to reduce the dimensionality of the dataset. As is seen in the heat map in (**Appendix 4**), there are certain variables that display a significantly high correlation whilst some other columns have a relevant and weak correlation. We selected 8 columns that displayed a relatively weaker correlation to be used for modeling. These final columns were:

Previously\_Insured', 'Vehicle\_Age\_2', 'Age\_Cat', 'Vehicle\_Damage', 'Annual\_Premium', 'Vintage', 'PSC', and 'Region\_28'.

# **Predictive Modeling**

## **Prepare for modeling:**

While we know that the y\_predict of our data set is "Response", we also need to determine the X features. As we see the result of the heat map and check all the correlations of all the variables and our target. We decided to take these columns as our X features:

-Previously\_Insured

-Vehicle\_Age\_2

- Age\_Cat

-Vehicle\_Damage

-Annual\_Premium

-Vintage

-PSC

-Region\_28.

## **Fixing the Imbalanced Dataset:**

However, data exploration revealed that this data set is imbalanced. Observations for one class outweigh the observations for the other. Therefore, it was necessary to fix this problem to avoid inaccurate results from models. Oversampling techniques can be used to fix this problem. Under sampling is one technique that fixes the imbalance in the dataset by removing observations from the majority class. However, this technique leads to loss of knowledge and useful information. Oversampling is another technique that works by creating and adding additional observation to the minority class. This technique however leads to the problem of overfitting. Given these issues, we chose to use SMOTE – which is the Synthetic Minority Oversampling Technique. This helped us increase the number of cases in a balanced way and fix the problem of imbalance.

## **Data Split:**

For our data set, we generated a data split between train and test

-train size:0.7

-test size:0.3

After we decided on the X features and dealt with the issue of imbalanced data, the data can be used to create models. In this part, we can try any model and we want to find the one that best fits our dataset. Our group tried two models to fit our dataset. One is logistic regression, and the other is a classification model.

# **Logistic regression model:**

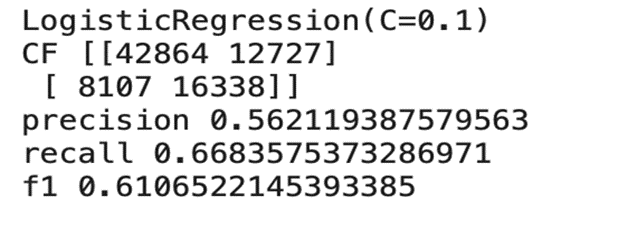
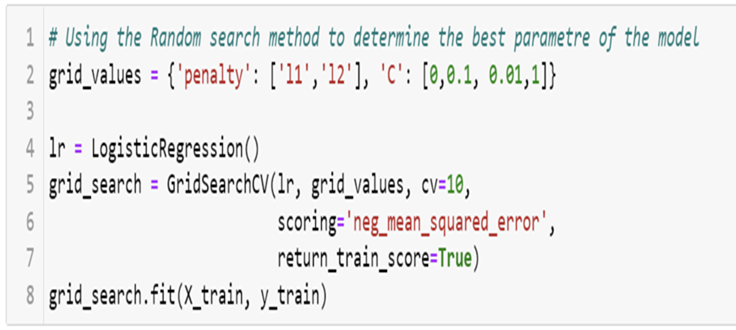
Predictive models built using this approach can make a positive difference in your business or organization. Because these models help you understand relationships and predict outcomes, you can act to improve decision-making. Binary is most useful when you want to model the event probability for a categorical response variable with two outcomes. In our analysis, the response variable is Yes - No type so the analysis can help assess whether the existing health insurance holder customer will buy the vehicle insurance.

Logistic regression provides us with the probability that the outcome will be yes or no. This provides us with the business opportunity to keep the threshold value of probability. For example, at 90% probability we can run our model predictions.

## **Lasso & Ridge Regression -**

Sometimes the model is overfitted because of having less data in the training data set that indicates that the fit line has high variance. The main idea behind Ridge regression is to find a new line that doesn’t fit the training data. The main advantage of Lasso and Ridge regression is that they apply higher penalties to the large coefficients in a regression equation.

In our analysis we used the penalty technique to get the best estimates for coefficients.



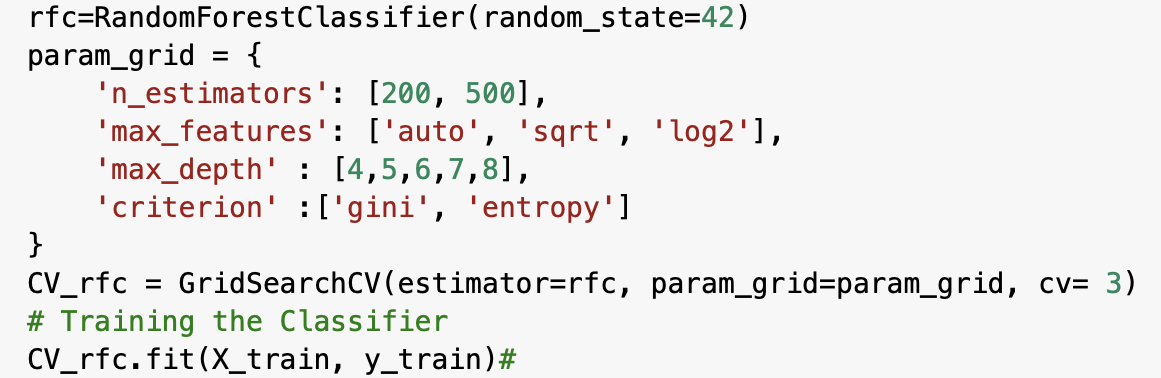
## **Evaluation Matrix -**

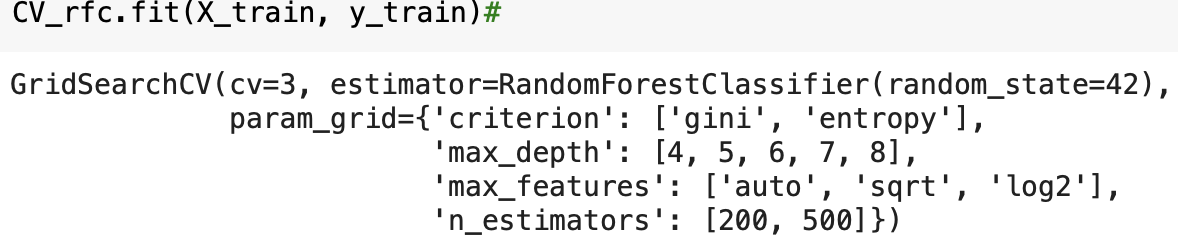
**F1-score** is one of the most important evaluation metrics in machine learning. It elegantly sums up the predictive performance of a model by combining two otherwise competing metrics — precision and recall.

For Logistic regression we got an F1 Score value of 0.6106, which meets the business requirement of 60%. However, we will perform classification trees as most data columns are categorical.

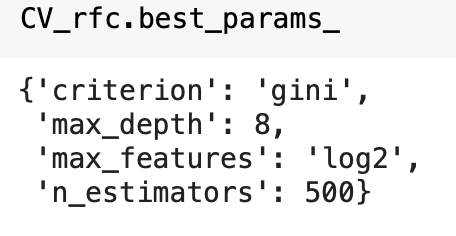
# **Classification Trees - Random Forest:**

-Creating the random forest classier by using grid search method

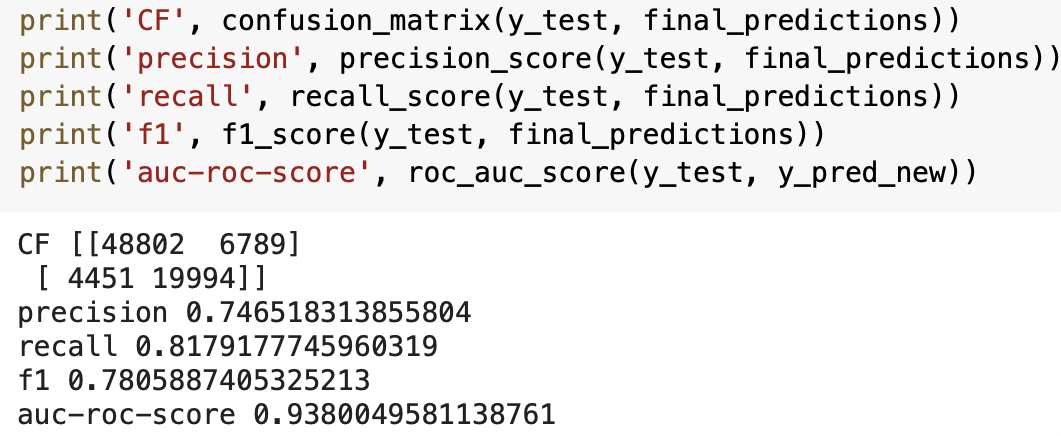


-Use training data to train the classifier and also to find out the best parameters of the model.

-Get the best parameters



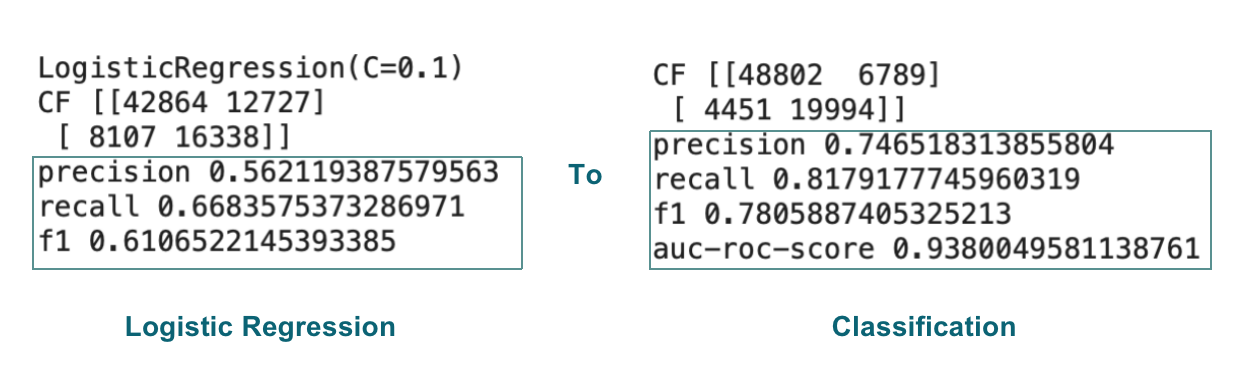
-Applying the training model to test data and get the confusion\_matrix, precision,recall and auc-roc-score of the model.



# **Findings**

## **Modeling conclusion:**

As we can see from the picture below, from logistic regression model to classification model, our precision, recall and f1 score are all increased, which means that the classification model does a better job in predicting the target. Therefore, if we only cared about the accuracy, the classification model would be better for our data set.

****

# **Highlights of Data Analysis:**

Through the data processing, we noticed that some values in column Vehicle\_age , Policy Sale Channel and Region Code have a significantly different relationship to our target. When vehicle age is greater than 2 years, when the Policy Sale Channel is 26 and 124, when the region code is 28. So, in order to deal with this situation, we take values which are vehicle age greater than 2 and region code is 28 out of their original columns and create two new columns to store them, then treat them as two dummy variables. For Policy Sale Channel, because we get a really high response when the sale channel is 26 and 124, so we regard them as 1, others as 0 to transform the data.

# **Correlations:**

From the heat map we make, we are able to find out all of the correlations of the target and other variables. We can conclude six variables that are relatively highly correlated with the target. Based on the correlation we can come up with some insights for business.

-Whether the vehicle is damaged before

-Whether the vehicle is previously injured

-The age of the vehicle

-The age of the customer

-The sale channels

-Whether the customer is from region 28

|  |  |
| --- | --- |
| Response & Vehicle\_Damage | 0.35 |
| Response &Previous\_Insured | -0.34 |
| Response &Vehicel\_Age | 0.22 |
| Response &Age\_Cat | 0.20 |
| Response &PSC | 0.18 |
| Response &Region\_28 | 0.12 |

# **Implications/recommendations for business**

Marketing teams can use our cross-sell predictions to grow their business and change business strategies accordingly. Our analysis suggested below recommendations for business to achieve growth -

1. Customers who got his/her vehicle damaged in the past are more likely to buy the insurance.
2. Customers who have never had their vehicle insured are unlikely to purchase automobile insurance.
3. Persons who own vehicles older than two years must pay a higher annual rate, which has resulted in a higher number of people in this category opting out of insurance. We need to adjust the amount somewhat so that folks in that category don't forego insurance. Vehicles that are at least 1-2 years old.
4. The number of days people work for the company has no bearing on the annual premium. As a result, we can change the premium policy in order for the insurance firm to attract additional consumers.
5. Overall, there are more negative responses than positive responses from customers, so we can assume that most of the insurance company's product offerings have degraded, or that customer after-sales service is poor, or that the product offering does not meet today's world customer needs, or that the insurance company lacks a marketing strategy. As a result, we can deduce that the insurance company's overall offers must improve, as well as the product's marketability, in order for those negative replies to be transformed into positive responses.
6. The Company can create specialized marketing strategies for customer Region Code 28. Also, most of the sales have been achieved using Policy Sales Channels 26 and 124.

# **References**

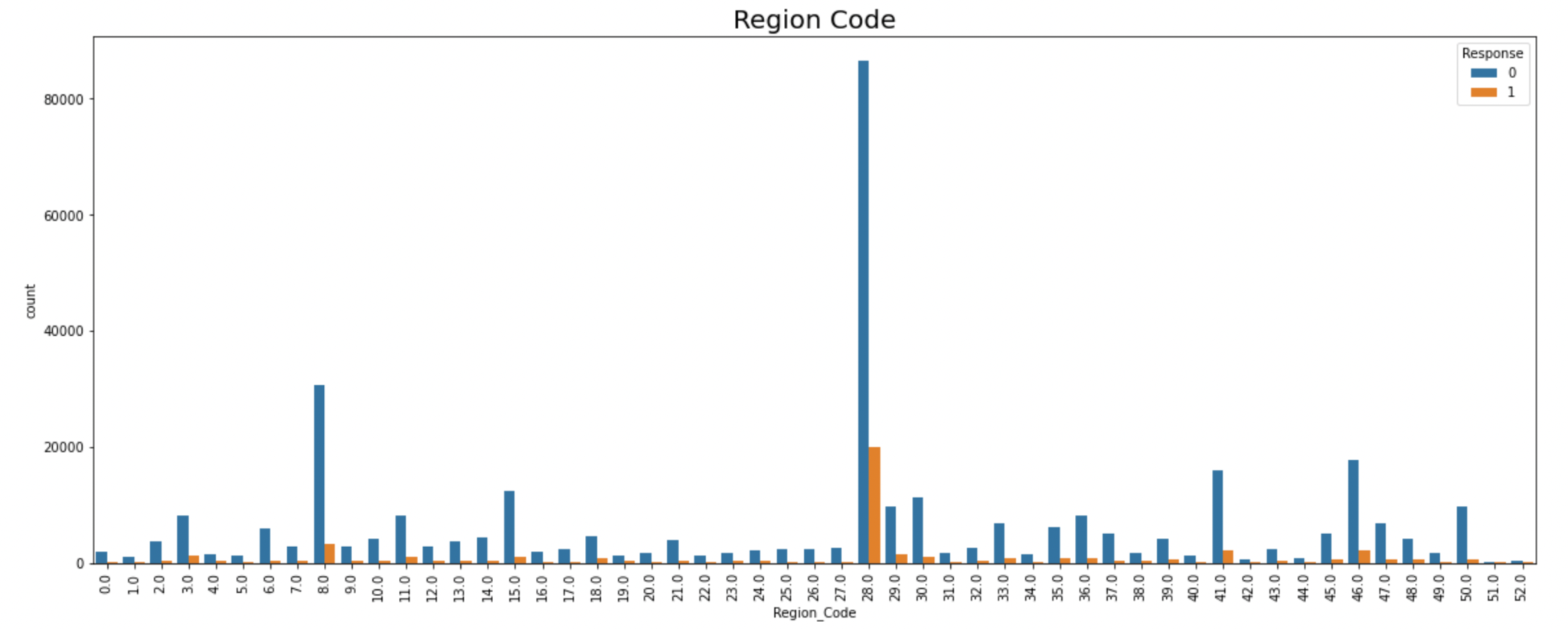
* We downloaded this dataset from a website with the URL - <https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction>
* Lecture modules and Class Slides
* Python Data Science Handbook - Jake VanderPlas
* We have taken some readings from –

<https://www.geeksforgeeks.org/binning-in-data-mining/>

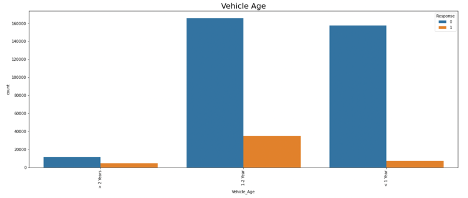
* AUC-ROC curve interpretation -<https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>
* Stackoverflow - <https://stackoverflow.com/>
* https://risk.lexisnexis.com/insights-resources/case-study/maximize-property-to-auto-cross-sell-campaign-performance

# **Appendix**

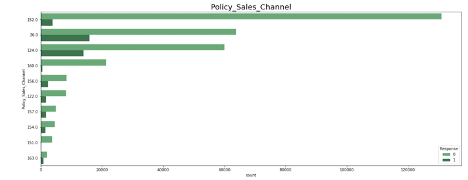
**1.1**



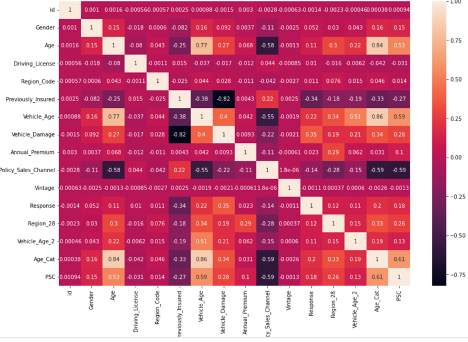
**1.2.**

****

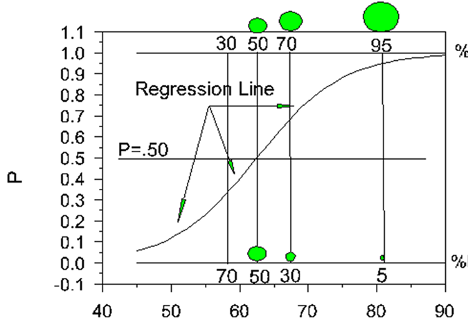
**1.3.**

****

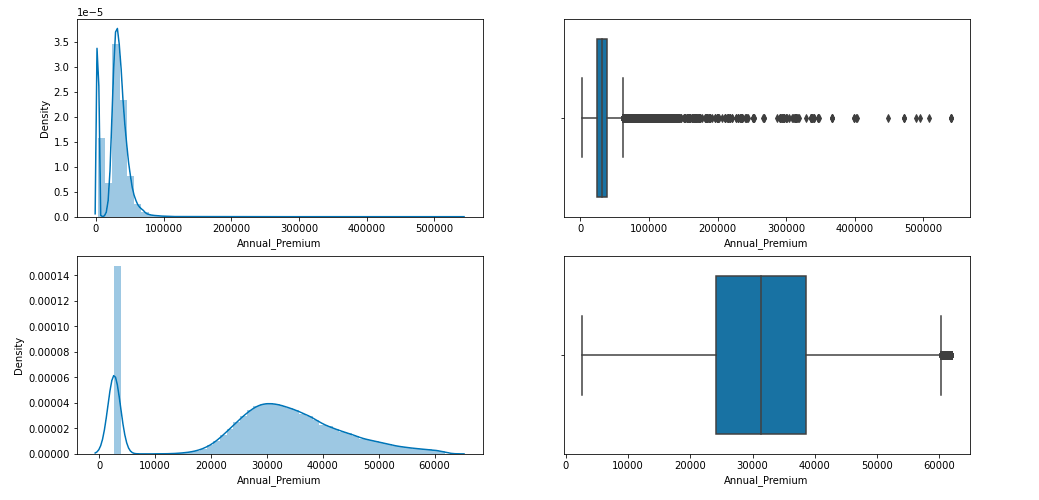
**1.4.**

****

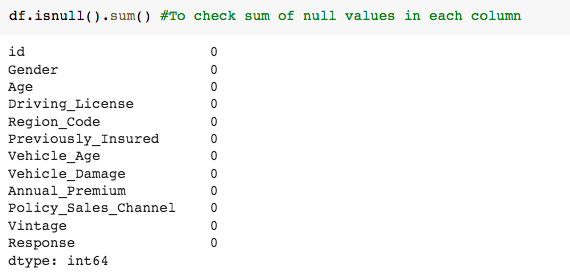
**1.5.**

****

**1.6.**

****

**1.7.**



**1.8.**

