IST-718 Project Proposal

Health Insurance Cross Sell Prediction

Group 6

**Grade: 97%**

Aatish Suman

Adheesh Phadnis

Sharvil Turbadkar

Snehal Ghatpande

**Objective**

The objective of the project is to build a model to predict whether the policyholders from the past year will also be interested in purchasing vehicle insurance provided by the company.

**Data Set Description**

**Overview/ Description**

The goal of the project is to build a model to predict whether a policyholder would be interested in vehicle insurance. It would be extremely helpful for the company because it can aid the company to plan its communication strategy to reach out to those customers and optimize its business model and revenue. To predict whether the customer would be interested in Vehicle insurance we are using data about the demographics (gender, age, region code type), vehicles (age, damage), policy (premium, sourcing channel), etc.

**Number of rows and columns**

The training data set consists of 381,109 rows and 12 columns while the test data set consists of 127,037 and 11 columns.

**Sample predictors**

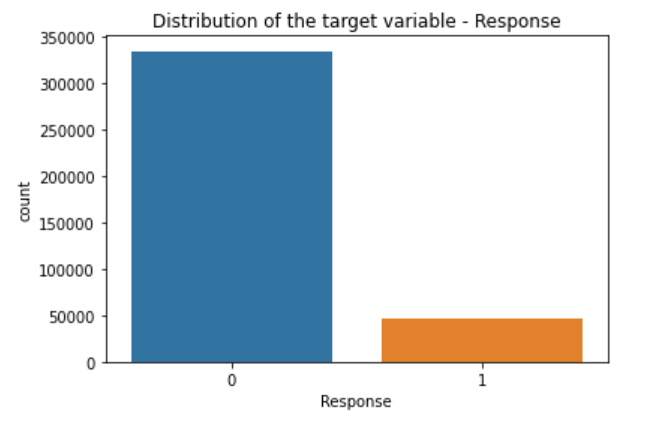
|  |  |
| --- | --- |
| **Variable** | **Definition** |
| id | Unique ID for the customer |
| Gender | Gender of the customer |
| Age | Age of the customer |
| Driving\_License | 0: Customer does not have DL, 1: Customer already has DL |
| Region\_Code | Unique code for the region of the customer |
| 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 |
| Policy\_Sales\_Channel | Anonymized Code for the channel of outreaching to the customer i.e. 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 |

**Link to the dataset** – <https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction>

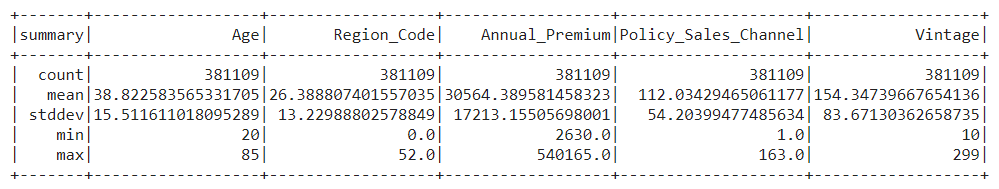
**Surprising fact about the data** – A few of the customers (0.2%) considered for vehicle insurance do not have a driving license.

**Preliminary Data Exploration**

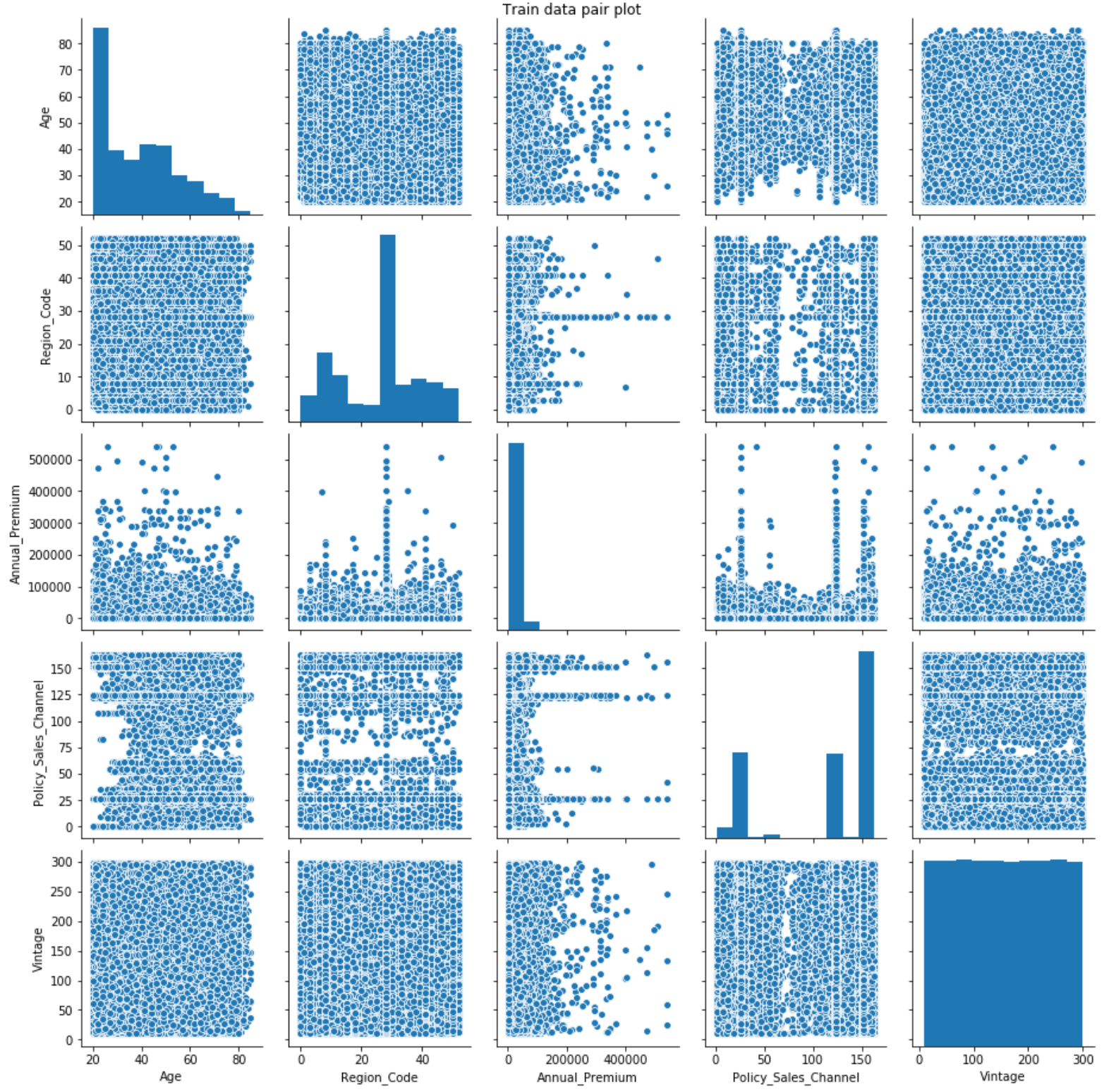
*Target variable* – The target variable (Response) is heavily imbalanced as can be seen from the bar plot below.



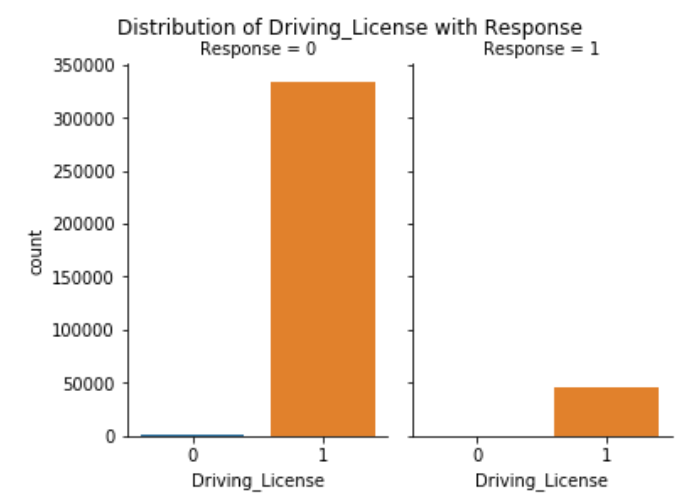
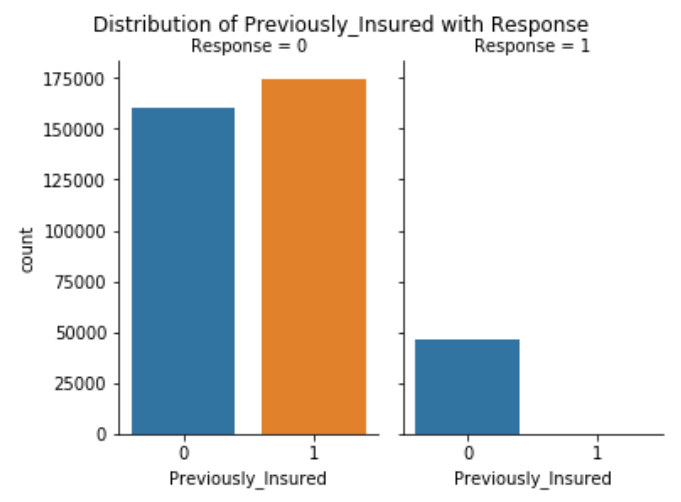
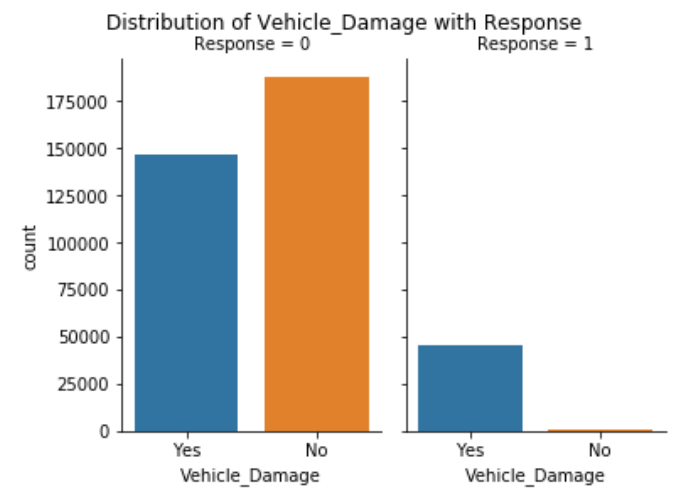
*Missing values* – The min, mean and max values for each numerical variable are comparable, which reduces the possibility of erroneous data. The data does not have null values. The scales of the variables are different, and scaling might be required for certain algorithms and to assess variable importance.



*Variable correlation* – There doesn't seem to be any correlation between the variables Age, Region\_Code, Annual\_Premium, Policy\_Sales\_Channel and Vintage as can be seen from the pair plots below.



*Interesting finding* – All interested customers (Response=1) have a driving license (Driving\_License=1) and already have a vehicle insurance (Previously\_Insured=1), and almost all of them have damaged their vehicle in the past (Vehicle\_Damaged=Yes) (shown in the categorical plot below). It shows that the 3 variables have high predictive power.

**Predictions**

* We plan to use PCA and linear regression to understand the relationship between different variables and estimate their predictive power for determining whether the customer is interested in purchasing vehicle insurance.
* We plan to use the Naïve Bayes-, SVM-, Random forests- and Gradient-boosted trees-based classifiers to predict whether a current policyholder would be interest in purchasing vehicle insurance from the company using customer demographics, vehicle properties, and policy.

**Inference**

* Which variables in the dataset play a critical role in predicting the target variable? Specifically, what characteristic of the customer makes him more likely to buy the vehicle insurance? This highlights the characteristics of the customer which narrows down the search strategy for the insurance company.
  + From a univariate analysis, we plan to gain information regarding the distributions of the individual features which might lead to feature engineering for specific algorithms, which in turn will lead to a better model for the company. From the univariate analysis we can also infer the relevance of each feature to target variable. For example, we can see if male/female tend more to subscribe for vehicle insurance etc.
  + From a multivariate analysis, we can plan to gain information about the correlation between the variables which is vital for dimensionality reduction, feature engineering etc. to build better models.
* Classify the customers using the supervised learning algorithms which concretely provides a model for the insurance company to classify the customers.

**Non-Spark Packages** – We will not be using any non-spark packages apart from the ones mentioned in the instructions.