**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Individual Project done By-:**  **NAME-** MONICA PATEL  **EMAIL –** [Monicapatel826@gmail.com](mailto:Monicapatel826@gmail.com)  **CONTRIBUTION -**  DATA CLEANING, EXPLORATORY DATA ANALYSIS, DATA VISUALISATION, REPORT GENERATION,PPT, TECHNICAL DOCUMENT,CLUSTERING , SUMMARY  DATA INSPECTION CHECKING THE NULL VALUES CHECKING DUPLICATES : DROPPING ID AND DRIVING LICENSE COLUMN. ▪EDA : GENDER WISE RESPONSE : RESPONSE WITH RESPECT TO PREVIOUSLY INSURED : RESPONSE WITH RESPECT TO VEHICLE DAMAGE : VEHICLE AGE COUNT : TOP 10 POLICY SALES CHANNEL : COUNT OF RESPONSE COLUMN ▪ CLASSIFICATION MODELS USED : LOGISTIC REGRESSION : NAÏVE BAYES : K-NEAREST NEIGHBORS : DECISION TREE |
| **Please paste the GitHub Repo link.**  Github Link:- https://github.com/1993monica |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**  PROBLEM STATEMENT: Cross-selling allows insurance companies to grow their bottom line without having to start from scratch. Similar to medical insurance, car insurance entails the customer paying a yearly premium to the insurance company in order for the insurance company to give the customer with compensation in the sad event that an accident involving the vehicle occurs. The objective is to forecast whether any potential new customers are likely to be interested in obtaining vehicle insurance from this company using the existing health and vehicle insurance customer data. By creating a model to predict if a customer would be interested in acquiring vehicle insurance, the business can then plan its communication strategy to reach out to those clients and maximise its business model and revenue. We have a dataset which contains information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. related to a person who is interested in vehicle insurance. We have 381109 data points available. APPROACH: • The first step includes loading of dataset and then inspecting the data through which we get to know the summary or description of data, shape and size of data, null value count, and duplicates values in the data and about the data types of column. • On the basis of univariate, bivariate, and multivariate analysis, we have carried out several visualisations. First, we performed a Univariate analysis since we needed to comprehend each feature or column's individual significance and the insights it would add to our study. Second, we used bivariate analysis to examine how one column or characteristic affects another, as well as the direction these discoveries may take us. Finally, we conducted a multivariate study to determine the effect of various factors on multicollinearity. • Next step involves visualization of data .In visualization we saw that dependent variable (i.e.response) is highly imbalanced and then used SMOTE technique to balance it. After having a look at the distribution of data we saw that Annual\_premium column have outliers. We convert Annual\_premium column to normal distribution by power transformer. • Following data visualisation, we utilize onehotencoder and label encoding to perform encoding, which converts categorical data to numerical data. We then performed feature  selection using VIF and removed variable Driving\_License because of high VIF value. We then divide the data by 80:20 using train test split. 20% for model testing and 80% for model training. • Then, various models are applied. We used Logistic Regression, Decision Tree, Random Forest Regression and XGBoost Classifier and then used Bayes search CV for hyperparameter tuning. |