**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| 1. **VISHAKHA KUMARI (**[**vishakhakumari0598@gmail.com**](mailto:vishakhakumari0598@gmail.com)**)**   **Contribution:**   * **Outlining project planning** * **Univariate analysis.** * **Data preparation & cleaning.** * **Technical documentation.**  1. **ABHINAV AKOTKAR (**[**akotkarabhinaw@yahoo.com**](mailto:akotkarabhinaw@yahoo.com)**)**   **Contribution:**   * **Decision making on dataset** * **Handling imbalance dataset** * **Fundamental step for analyzing dataset.** * **Project summary**  1. **DURGESH SHUKLA (**[**@yahoo.com**](mailto:akotkarabhinaw@yahoo.com)**)**   **Contribution:**   * **Correlation heat map** * **Linear regression** * **Data exploration.** * **Power point presentation of model**  1. **SHIVAM (shivamself1997@gmail.com)**   **Contribution:**   * **Analysis of the data based on visualization.** * **Model selection.** * **Location visualization on map** * **Maintaining colab notebook** |
| **Please paste the GitHub Repo link @ drive link** |
| Github Link:- https://github.com/shivam0070/Health-Insurance-Cross-Sell-Prediction  Drive Link:- https://drive.google.com/drive/folders/1HMX-sj4gtwDfrLn0Dkd0TshgXSvLSbfS |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Health Insurance Cross Sell Prediction is a predicting Health Insurance Owners' who will be interested in Vehicle Insurance |
| 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 customers from past year will also be interested in Vehicle Insurance provided by the company. Our objective is to 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 optimize its business model and revenue. Database contain 381109 entries and total 12 columns contain -: id, gender, Age,  Driving License, Region Code, Previously Insured, Vehicle Age, Vehicle Damage  Annual Premium, Policy Sales Channel, Vintage, Response by using Pandas.  First step is to Importing some important libraries , preparing the problems and summarized data  Analysis and Visualization of dataset by using matplotlib.  The graphicalrepresentation between Response and Count shows The data is highly imbalanced.  Then graphicalrepresentation between Gender and Count shows gender variable in the dataset is almost equally distributed and Male category is slightly greater than that of female and chances of buying the insurance is also little high.  We notice between Age Vs Response that  Young people below 30 are not interested in vehicle insurance. Reasons could be lack of experience, less maturity level and they don't have expensive vehicles yet.  People aged between 30-60 are more likely to be interested.  From the boxplot we can see that there no outlier in the data.  Driving License vs count shows Customers who are interested in Vehicle Insurance almost all have driving license.  Previously Insured Vs Response showsCustomer who are not previously insured are likely to be interested.  Annual Premium From the distribution plot we can infer that the annual premium variable is right skewed From the boxplot we can observe lot of outliers in the variable  After that using correlation analysis we notice that Target variable is not much affected by Vintage variable. We can drop least correlated variable.  Feature Selection technique shows We can remove less important features from the data set by using tree based classifier.  Handling Imbalanced data analyzing noticed that When observation in one class is higher than the observation in other classes then there exists a class imbalance. We can clearly see that there is a huge difference between the data set. Solving this issue we use resampling technique. After that Model Selection analyzed took placed to start selection of models as -: Logistic Regression, Random Forest, XGBClassifier. By using all this analyzing and visualization we find the best conclusion for dataset. |