**Capstone Project (Health Insurance Cross Sell) Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| **Name:** Indugopal Maity **Email-Id:** [ig.maity@gmail.com](mailto:ig.maity@gmail.com)  **Contribution:**   1. Data Wrangling    * 1. Unique Id, Gender, Age of Customer      2. Driving License available or not, Region Code, Previously Insured or not, Vehicle Age, Annual Premium, Policy Hold Channel, Vintage and Responses of the targeted customers. 2. Found Missing Values. 3. Calculated relationship between “Interested” and “Not Interested” for a Vehicle Insurance. 4. Calculated relationship between “Response Count” and “Gender”. 5. Calculated relationship between “Response Count” and “Age”. 6. Calculated relationship between “Response Count” and “Driving License”. 7. Calculated relationship between “Response Count” and “Previously Insured or not”. 8. Calculated relationship between “Response Count” and “Vehicle Age”. 9. Calculated relationship between “Response Count” and “Annual Premium”. 10. Plotted correlation between variables. 11. Checked Duplicate rows. 12. Plotted Feature Selection 13. Handled Imbalance Data 14. Split data into train and test data. 15. Plotted different ML Models (Logistic Regression, RandomForest Classifier and XGBClassifier). 16. Drawn several ROC curves with respect to different ML Models. 17. Compared the ML models to check which performs the best out of three ML Models. 18. Best on the different ML Model wrote the conclusion. |
| **Please paste the GitHub Repo link.**  **https://github.com/indugopal1991/Health-Insurance-Cross-Sell/blob/main/Indugopal\_HEALTH\_INSURANCE\_CROSS\_SELL\_PREDICTION.ipynb** |
| Github Link:- https://github.com/indugopal1991 |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**  As the first step, performed **“DATA Wrangling”** over the raw data, **“TRAIN HEALTH INSURANCE CROSS SELL PREDICTION.csv**”. We can see that there is a combination of “**381109”** rows and “**12”** columns. The chart is consisting of several columns like **“Id”**, **“Gender”**, **“Age”**, **“Driving\_License”**, **“Region\_Code”**, **“Previously\_Insured”**, **“Vehicle\_Age”**, **“Vehicle\_Damage”**, **“Annual\_Premium”**, **“Policy\_Sales\_Channel”**, **“Vintage”** and **“Response”**.  Farther we have divided the complete project into seven different parts as per the **“Target variable”**, **“Features v/s Responses Count(EDA)”**, **“Correlation Analysis”**, **“Checking Duplicate Rows”**, **“Feature Importance”,** **“Model Selection”** and **“Comparing Models with points”**  In the first part, we tried to find out **Target Variable**, and by this data we figured out data is highly imbalanced, like most of the responses don’t like to have a Vehicle Insurance.  In the second part, we tried to find out the relation between **“Responses Count”** to several **“Features”** such as **“Gender”, “Age”, “Driving\_License”, “Previously\_Insured”, “Vehicle Age”** and **“Annual Premium”**.  In the third part, we have checked **“Correlation Analysis”**, to check which are **positively correlated** and which are **negatively correlated**.  In the fourth part, we have checked **“Duplicate data rows”**. And we can see that there are **no duplicate rows** in the dataset.  In the fifth part, we have seen which **features** have the least importance for **Vehicle Insurance** and those are **“Driving\_License”** and **“Gender”**.  In the sixth part, we have used different **“Regression Model”** such that **“Logistic Regression”, “RandomForest Classifier”,** and **“XGBClassifier”** along with **“ROC Curve”**. After using different types of regression model, we can see that **“RandomForest Classifier”** is the best as per the character of **“ROC Curve”**.  In the seventh and last part, we have compared **“regression Models”** with respect to different points such that **“Accuracy”, “Recall”, “Precision”, “fi\_score”** and **“ROC\_AOC”** and we can say that **“RandomForest Classifier”** is working best as per the points we have checked**.**  So, by the analysis we have figured out certain things, those are listed below:   1. Customers between **“30-60”** are more **likely** to **buy Insurance**. 2. Customers with **Driving License** have a higher **chance** of **buying Insurance**. 3. Customers with **Vehicle\_Damage** are **likely** to **buy Insurance.** 4. The variables such that Age, Previously\_Insured and Annual Premium are more affecting the target variable. 5. Comparing **“ROC Curve”** we can see that **“RandomForest Classifier”** performs better because it curves closer to the top-left corner, which indicates a better performance. |
|  |