**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| | Names | Email ID | Contribution | | --- | --- | --- | | Ankush Kumar  (Team Leader) | dsankushkumar@gmail.com | Code:  Different Variate Analysis and correlation.  Modeling Hyperparameter tuning  Conclusion and Summary  Power Point Presentation Preparation | | Nayan Kumar Jha | nayan8625@gmail.com | Code:  Data Insights  Data Visualization  Feature Transformation  Power Point Presentation Preparation | | Pinky Thakur | thakurpinky1896@gmail.com | Code:  Data Information  Data Wrangling  Data Preprocessing  Power Point Presentation Preparation | |
| **Please paste the GitHub Repo link.** |
| GitHub Link: - <https://github.com/dsankush/Health_Insurance_Cross_Sell_Prediction> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**  In exchange for the payment of a predetermined premium, a firm promises to guarantee reimbursement in the case of a certain loss, damage, disease, or death under the terms of an insurance policy.  A premium is the amount of money that the client must pay to an insurance company on a regular basis in exchange for this promise. Car insurance operates in the same way that medical insurance does in that consumers must pay an annual premium to the insurance provider firm in order to be rewarded (referred to as the "sum assured") in the event that their vehicle is at fault in an accident.  a client who already has insurance would be interested in purchasing vehicle insurance by developing a model to anticipate if a consumer would be interested in Vehicle Insurance, the company can then plan its communication strategy to reach out to those customers and optimize 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  There was not any null value and also, no duplicate data was present in our data.  We have performed different types of visualization on the basis of Univariate, Bivariate and multivariate analysis. Firstly, we have conducted Univariate analysis as we also need to understand various features/columns individually and what kind of importance and insights they bring for our analysis. Secondly, we have performed Bivariate analysis so that we can analyze the impact of one column/feature to another feature and where these insights lead us. At last, we performed Multivariate analysis in which we came to know the impact of multiple features.  Area code 28 has the most health insurance clients and region code 51 has the fewest.  Almost no effect on insurance buy if a person is already having insurance or not. vehicle age between 1-2 years has the greatest number of counts and vehicle age less than 2 years has the least count. Almost equal count for both the Vehicle property based on damage. The average age of customers with auto insurance is 34.5 while the average age of customers without auto insurance is 42.4. Age range of vehicles Customers with newer cars are less likely to be interested in purchasing auto insurance.  Hyper parameter tuning, the False Negative rate of Logistic Regression is 11970, which is the highest among all models, and the False Positive rate of Random Forest is 9184, which is also high. When we compare the false positive rates of KNN and XGBOOST, we get 3080 and 5001 respectively, which is lower than the Logistic and Random Forest models but higher than the Decision Tree Decision Tree Classifier, which has the lowest False Positive Rate of 1818 of all the models. However, if we consider good recall, accuracy, and a low false positive rate, KNN does a great job with (96 percent recall, 83 percent accuracy, and a 3080 false positive rate) if all parameters are considered, our final model should be KNN Classifier However, if we are only concerned with Recall Rate, we should use Decision Tree Regressor. As we improve the KNN Classifier Model with more hyper parameter tuning, we will get more desired results that will be more useful. |
|  |