Capstone Project Submission

Team Member's Name, Email and Contribution:

1. **Suvir kapse**

Email- suvirkapse@gmail.com

- Exploratory Data Analysis:
  - Analyzing responses based on gender.
  - Age, Previously Insured, Vehicle age, Region code Vs Response
  - Checking distributions.
- Imbalance techniques
- Preparation and Model making
- Model training
  - Logistic Regression
  - Decision Tree
  - Random Forest Model
  - XGB Classifier Model
- Hyperparameter tuning for top 2 models
  1. For Random Forest Classifier
  2. Xgboost classifier

2. **Ajinkya Dakhale**

Email- Ajinkya.dakhale2408@gmail.com

- Data inspection
- Exploratory Data Analysis
  - Checking distribution of features.
  - Checking relation of target feature with independent feature.
  - Used multiple graphs and did analysis on dataset.
- Feature Engineering
  - Checking null values.
  - Handling outliers.
  - Encoding categorical features.
- Feature Selection
  - Variance Threshold
- Model training
  - Logistic Regression
  - Decision Tree
  - Random Forest Model
  - XGB Classifier Model
- Imbalance Techniques
- Technical Documentation

3. **Harshjyot Singh**

Email- hs9158695878@gmail.com

- Data Cleaning
- Outlier handling
- Checked distribution of numerical feature
- Checked correlation
- Encoding for categorical features
- Features Selection
- Preparation and Model making
- Model training
  - Logistic Regression
  - Decision Tree
  - Random Forest Model

        - XGB Classifier Model
    - Imbalance Techniques
    - Set up Roc and Precision recall curve for best model
    - Conclusion

Please paste the GitHub Repo link.

Github Link:- https://github.com/SuvirKapse/HEALTH-INSURANCE-CROSS-SELL-PREDICTION

Problem Statement:

Our 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 policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalized in that year, the insurance provider company will bear the cost of hospitalization etc. for up to Rs. 200,000. Now if you are wondering how can company bear such high hospitalization cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2-3) would get hospitalized that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called 'sum assured') to the customer.

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.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

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.

Conclusion:

We used different type of algorithms to train our model like, Logistic Regression, Random Forest model, Decision tree and XGB Classifier. And Also we tuned the parameters of XGB Classifier and Random Forest model Comparing the model on the basis of precision,recall, accuracy ,F1 score we can see that the XGBClassifier model performs better.

- KEY POINTS

  - The male customers own slightly more vehicles and they are more tend to buy insurance in comparison to their female counterparts.
  - The people not interested is 87 % as compared to the interested once.
  - Similarly, the customers who have driving licences will option for insurance instead of those who don’t have it.
  - From the plot, we can conclude that if the vehicle has been damaged previously then the customer will be more interested in buying the insurance as they know the cost.
  - The customers with vehicle age lesser than the 2 years are more tend to buy insurance as compared other.
  - 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.
  - We found out that customer who already have vehicle insurance almost have no interest in another vehicle insurance.
  - People aged between 30-60 are more likely to be interested.
  - More than half of the customers does not have a vehicle insurance .
  - We can conclude that if the vehicle has been damaged previously then the customer will be more interested in buying the insurance as they know the cost.