**JOB-A-THON**

Logo, company name

Description automatically generated

**Approach:**

**EDA:**

Exploratory Data Analysis is always the step of any machine learning lifecycle. It involves different steps like Univariate Analysis, Bi-Variate analysis, checking for missing values, outliers, missing value imputation etc. Steps followed and some salient observations were as follows,

1. **Univariate Analysis:**
2. 3 of the variables had missing data, namely Health Indicator, Holding\_Policy\_Duration and Holding\_Policy\_Type. It was also found that the missing values in Holding\_Policy\_Duration and Holding\_Policy\_Type were correlated.
3. It was found that the Response variable was not balanced. 75% of the respondents had not opted for the health insurance and only 25% had opted for health insurance.
4. It was also found that Region\_Code variable was a high cardinality categorical variable. It was also found that some of the Region\_Code values present in the training data were not present in the test data.
5. It was found that most of the recommended insurance types were Individual in nature and few were joint health insurances.
6. It was found that for individual health insurance, both the lower and upper values were same.
7. It was also found that most of the customers were not married to each other.
8. It was also observed that health of around 20% of the customers was missing.
9. During the initial exploration of the data, it was found that if a customer had missing data about Holding\_Policy\_Duration, the same customer had missing data about Holding\_Policy\_Type variable. It may be because the customer did not have any existing insurance with FinMan financial services.
10. It was also found that there are a variety of Health Insurance policies that are recommended to the customers, 22 to be precise. And Policy type 22 seemed to be most popular among the customers.
11. It was found that the distribution of Policy Premiums was same in both test and train data, and it was also found that the distribution was slightly positively skewed. Although there were some high premiums, they were not high enough to cause alarm.
12. **Bi-Variate Analysis:**
13. During the Bi-Variate analysis, it was found that the response rate was almost the same for all the City\_Code values and it was hovering around 25%. i.e., only 25% of the customers had responded positively. And this kind of response rate was observed with other variables like Accomodation\_Type, Reco\_Insurance\_Type, Is\_Spouse, Health Indicator variables.
14. It was found that Recommended Policy Category values of 1 and 4 had the least response rate and Recommended Policy Category value 15 had the highest response rate.
15. It was also found that the distribution of Policy Premiums for both Response values was almost the same, and it was slightly positively skewed.
16. It was also found that as the upper age increased, the policy premiums also increased, just as it is expected. This relation did not hold true with the lower age variable.
17. **Correlation:**

It was found most of the dependent variables were not correlated with the independent variable, the correlation was not strong enough in either direction. Only the Reco\_Policy\_Cat variable had a better correlation compared to the other variables. It was also found that Lower age and Upper age were highly correlated with each other. And as we had seen earlier, Upper Age variable was highly correlated with the policy premium variable. Below is the diagram representing the correlation,

Chart

Description automatically generated

1. **Missing Value Treatment & Feature Generation:**
2. Separate indicator variables were created to show the missing values in each column.
3. Holding Policy duration variable was coded as of object data type because of the value 14+, it was converted to integer data type and missing values were replaced with 0.
4. Similarly, missing values in Holding policy type variable were replaced with 0.
5. It should also be noted that some other large values were tried as the missing value replacements, but 0 proved to be more efficient.
6. Age, Policy Premiums were binned and used as features to build our model.
7. A lot of Grouped features were generated based on different categorical columns in the data, some of them include Sum, Min, Max, Mean, Median, values of continuous variable like policy premium grouped by categorical data.
8. Rank features were also generated which proved to be useful features. Ranking basically ranks the continuous variable by their value, there can be many methods used in groupby ranking such as min, max, average, first etc.
9. Count features and unique values based on other categorical features were also generated to be used for modelling.
10. Features were also generated based on other findings while exploring the data. It was found that the response rate of customers for whom individual insurance was suggested and if they were not married, they had higher chance of opting for health insurance. Many such features based on observations were tried and tested.
11. Label Encoding was used to be encode the categorical features.

**Modelling:**

Modelling always comes second to feature engineering and exploratory data analysis, having said that, it plays an important role in the entire predictive modelling purpose.

1. Light GBM was used for modelling purpose. Light GBM is more efficient and faster as compared to other models such as XGBoost or CatBoost, XGBoost didn’t perform any better than the LightGBMmodel.
2. For validation purpose, a 10-fold StratifiedKFold CV model was used. Stratified Cross Validation technique was used as the data is skewed and with normal KFold CV, the data distribution in the folds is not the same.
3. Up Sampling, Down Sampling and Synthetic Samples were generated to offset the class imbalance, but none of these methods were helpful in the end.
4. Hyperparameter tuning was extensively performed tuning the values of learning rate, max depth, subsample, column samples, etc.
5. Regularization was also performed so that the model does not overfit on the training data. The maximum roc\_auc score on the public leaderboard reached was 0.715 with a 10Fold Stratified Cross Validation model using LightGBM.

**Room for Improvement:**

There is always room for improvement in the field of machine learning.

1. Stacking, Ensembling should have improved the model performance.
2. Generating more features based on domain knowledge always helps the cause.