Project 1

Loan Prediction

1. **Problem Statements:**

Housing Finance Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers

.Aim:- To build a predictive model to predict whether a loan would be approved or declined

1. **Data Gathering:**

They have given the access of their database. Database is MySQL. Data is having 66 features. During POC data size is around 1.5 lakhs. After the POC we received the around 10 lakh data.

We split the data for training and testing purpose. We keep 80% for training and 20% for testing.

1. **EDA: Exploratory Data Analysis.**

After the data gathering, we moved to EDA part.

Then we separate out the categorical feature and numerical features for analysis purpose. There are 42 categorical features and 24 numerical features. To understand each feature, we did analysis using visualization.

**Categorical features names:**

1. policy type
2. Gender
3. Education
4. Area code
5. Marital status
6. Employment status
7. Renew offer type
8. Vehicle class
9. Vehicle size

**Numerical features name:**

1. Monthly premium
2. No of policies
3. Income
4. Total claim amount
5. No of open complaints
6. Application Id
7. Dealer ID

**EDA for categorical data type features:**

1. Number classes in each feature and its count using counter plot.
2. Missing value

**EDA for numerical data type features:**

1. Outliers use boxplot.
2. Normality using matplotlib and seaborn library we found that data is not normally distributed.
3. Missing values.
4. Bivariate analysis
5. **Feature engineering:**

After completion of EDA part and discussion with customer we moved to feature engineering.

**Missing value:**

We handle the missing by knn imputer. In this method we can consider gender column as target column and remaining two or more column as independent features to predict the NaN values.

**For categorical:**

We replace the missing value by Mode for categorical features like employment status.

**For numerical:**

We have replaced the missing value with mean of the columns like Age.

zero for no of complaints.

**Label encoding / one hot encoding:**

We did the label encoding (ordinal Encoder) on categorical features like marital status, gender and one hot encoding on policy type, location code after doing this data size increase to 105

**Handle the outliers:**

We did the transformation i. e. normalization on features like income, monthly premium.

1. **Feature selection:**

After completion of feature engineering. We moved to feature selection.

Part 1:

At the initial we dropped features which contains less information. Like those having unique value, dates, timestamp, application no., dealer id.

Part 2:

We have used below feature selection techniques:

1. Corr test :Kendall test (Num vs Cat)
2. Chi2 test (Cat vs Cat)
3. Kruskal Wallis H test (cat vs cat)
4. Mann Whitney U test (Num vs Cat)
5. Forward selection method (Num vs Cat)

We got the best features by forward selection method. We have finalize 25 features for final model training.

1. **Model training:**

We selected the below algorithm for building the model

1. Linear regression:

due to less chances of over-fitting. But it is sensitive outliers, data requires linear.

1. Decision Tree Regressor:

It is less sensitive to outliers; no feature scaling requires. But it mostly overfits.

1. Random Forest Regressor:

It has less chances of overfitting. It is having OOB error

1. XG-boost Regressor:

It also less overfits.

To confirm our build model should work perform so we did the cross validation after every model. We also did the hyper parameter tuning on decision tree, random forest and XG-boost Regressor

1. **Evaluation:**

During every model training we compute the RMSE, accuracy and R2-score. So, we found that XG boost has good performance than other.

We also tried adjusted R2-score.At the final RMSE is 0.2166 and R2-Score is 0.96

1. **Deployment.:**

Our deops team deployed our project on customer server.

Theory:

Customer lifetime value (CLV) is a business metric that measures how much a business can plan to earn from the average customer over the course of the relationship.

Customer lifetime value (CLV) is a measure of the total income a business can expect to bring in from a typical customer for as long as that person or account remains a client.

When measuring CLV, it’s best to look at the total average revenue generated by a customer and the total average profit. Each provides important insights into how customers interact with your business and if your overall marketing plan is working as expected.

**CLV =** Average Transaction Size x Number of Transactions x Retention Period