**Project**

**Telco Customer Churn Prediction:**

Telecom customer churn refers to the phenomenon where subscribers discontinue or switch their services from one telecommunications provider to another. Churn is a critical metric for telecom companies, as it directly impacts revenue and profitability.

Several factors contribute to customer churn in the telecom industry, including dissatisfaction with service quality, pricing issues, competition, changing customer needs etc..

Objective of this project is to predict the churning of any customer based on the given dataset.

To achieve the objective, a classification filter to be created which can determine the customer churn from shared telecom dataset.

Entire Project is divided into various steps:

1. Data Preparation
2. EDA
3. Feature Engineering
4. Predictive Modelling
5. Algorithm Results
6. Conclusion

1. Data Preparation:

Given csv format data is loaded into the notebook for analysis.

2. EDA:

i. It is found that some of the columns are not properly named, these were rectified.

ii. Some of the values in TotalCharges variables were blank. This was rectified after changing the datatype from object to float.

iii. Data Imbalance is observed for dependent variable Churn which can be considered during modelling to resample using SMOTE synthesis.

Through various plots and graphs, below analysis done for categorical variables and concluded for modelling.

iv. Churn Analysis Based on Services Opted:

Higher Churning for those who are having Services like Internet Service over OpticalFibre.

Higher Churning for those who doesn't opted for services like Online security, Online Backup, Device protection.

Higher Churning is observed for those where there was no Tech support.

v. Churn Analysis Based on mode of Customer Account Operation:

Higher Churning rates are for those who are on Monthly Contract.

Higher Churning observed who opted for Electronic mode of Payment Method.

vi. Churn Analysis Based on mode of Demographic info of customers:

Higher Churning rates are observed for those who doesn't have Dependent or Partner Connection.

vii. TotalCharges and Tenure are highly Correlated, signifies both variable relation with Churn should be Similar and can be removed if require.

Similarly, below is concluded for Numerical variables:

viii. Higher Churn rate when Monthly Charges are high.

ix. Senior Citizen Churn rates higher than Lower Age Population.

x. Higher the Tenure of stay, Lower the Churn Rate.

3. Feature Engineering:

i. All Categorical variables are converted into Numerical through label Encoding.

ii. It is require to scale all numerical variables e.g: Tenure,MonthlyCharges,TotalCharges to be scaled in same range. That’s why using MinMax Scaler all numerical variables are converted in the range of 0-1.

After doing EDA and Feature Engineering and taking the correlation among all the variables, below observation deduced:

As deduced earlier with EDA which is evident here as well:

1. TotalCharges and Tenure are highly coorelated variables.
2. TotalCharges and MonthlyCharges are positively coorelated.
3. Tenure and Contract term are positively coorelated.
4. Churn rate is negatively correlated with different services opted (e.g: OnlineSecurity,OnlineBackup,DeviceProtection,TechSupport,Contract Term).
5. Churn rate is also negatively coorelated with Tenure and TotalCharges as an obvious and expected.

4. Modelling:

Below classifier algorithms were tried to achieve the objective:

1. Logistic Regression
2. Random Forest
3. Boosting Technique: AdaBooster
4. Cross Validation Technique: Stratified K-fold CVC

Also as observed during EDA, there is class imbalance of Churn variable, so modelling was also done by handling this class imbalance using oversampling method using SMOTE.

Models tried after class imbalance were:

1. Logistic Regression
2. Random Forest
3. Boosting Technique: AdaBooster

5. Algorithm Results:

Accuracy Results:

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic | 81.2% |
| Random Forest | 79.5% |
| AdaBooster | 78.9% |
| K-Fold CVC | 71.8% |
| After handling Class Imbalance | |
| Logistic | 77.4% |
| AdaBooster | 75.5% |

Confusion Matrix and ROC-AUC Curve is used to conclude the model.

6. Conclusion:

After trying with multiple modelling, it is observed that Logistic Regression is giving good accuracy of around 81% and we will select this model for the project.