Telecom Customer churn prediction using Machine Learning.

Domain Background

Customer churn is basically when a customer chooses to end your services and switch to another provider.

Churn rate is defined as the number of individuals that move out of a particular group over a period of time. Any business is dependent on its customers for it's revenue especially in a service based industry like telecom, banks. It becomes extremely vital for banks to retain customers in order to benefit from their collective funds in the bank. Determining factors which can explain why a particular customer leaves you can enable banks/telecom to take corrective measures or offer better services to the customers. Losing clients, especially for a company whose revenue is directly dependent on subscription based models is very costly. Telecom companies today face stiff competition and have to rival cheaper offers and better services provided by a rival company.

It is extremely difficult to bring back those customers once they make the switch. I myself have switched from one company to another, and refuse to go back to my previous provider, in spite of multiple discounted offers from them once I made the switch, simply because of the following factor- poor network service and equally high rates.

Reference papers:

- A review and analysis of churn prediction methods for customer retention in telecom industries by Ammara Ahmed and D Maheshwari.
- **Predicting Customer Churn in Telecom Industry using Multilayer Preceptron Neural Networks: Modeling and Analysis by Omar Adwan, Hossam Faris, Khalid Jaradat and Osama Harfoushi.**

Problem Statement:

The problem to be solved is to predict which customer if and when a customer is likely to churn so that companies can offer better incentives to those customers. We will be trying out both ML algorithms like Logistic Regression, Decision Trees and Artificial Neural Networks as well to Predict Customer churn. This is a classification problem which needs us to classify data into 2 categories, in 0 and 1, 0 indicating a customer is not likely to churn and 1 indicating the customer is likely to churn.

Data Set:

IBM Watson Telecom customer churn Dataset
https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/
The name of the Data set is WA_Fn UseC_ Telco Customer Churn.csv.

This data set contains 7043 rows and 21 columns. Currently the dataset doesnot seem to have an imbalanced dataset.

Columns:

Churn: the column we're trying to predict. This has a value of 0 or 1. This represents which customer left churned within the last month.

Service Information: The types of services provided to customers-phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movie. They contain Yes or No as values.

Demographic information

Gender: Male or Female

Senior Citizen
Partners: Yes or No Values.
Dependents: Whether they have anyone dependent on them financially. Yes or No values.
Customer account information.
Customer id: Unique id of customer of type char
Tenure: how long they've been a customer of this company.
Payment method: Contains values like Electronic check, mailed check, bank transfer
Paperless Billing: Values of Yes or No. Indicator of whether customer has opted for paperless billing.
Monthly charges: Charge per month. Numeric values.
Total charges: Contains numeric values. Indicates charges over the contract period.
Contract: Indicates whether the customer has a month to month contract, or a yearly contract.
Inputs- All columns except churn column would be the input.
Output- the column Churn is the column we're trying to predict.

Solution.

The column that we are going to predict has a label so we can try out Supervised learning algorithms like Logistic Regression, Random forest, and try boosting algorithms like XGBoost and Light GBM. The project will also develop a artificial neural network-ANN(feed forward neural network) to be able to classify where a user falls. The main focus of this capstone however is to try and develop an artificial neural network for this problem.

Evaluation Metrics:

Accuracy of the model will be used as an evaluation metric. Precision/recall are other metrics. The ROC curve can also be used.

Bench Mark Model.

As the intention is to explore neural networks and XGBoost , Logistic Regression or a random forest could be taken up as a benchmark model here.

Project Design.

Step 1:
Exploratory Data Analysis:

- Loading necessary libraries
- Statistical summary, analysis and visualization.

Step 2:
Data Preprocessing.

- Identify feature and label columns. Drop the label column from the dataset.
- Cleaning up the data, dealing with missing values.
- Determine feature relevance.
- Use imputing or replacing with average values for missing data.
- Use techniques like One hot Label encoding techniques if needed, to convert categorical variables to zeroes and ones.
- Splitting into training, validation and test sets.
- Standardizing or normalizing data if necessary.

Step 3:
Training your model/neural network.

- Build the models like Logistic regression
- Build an artificial neural network.
- Train your model.

Step 4
Test performance.

- Use validation and testing data sets to test performance.

- Determine accuracy of the model.
- Use techniques like GridSearchCV to select best model.

Step 5

Model tuning.

- Improve model performance by tuning hyper paramters.

Step 6.

Final conclusion and closing comments or observations.