Churn Prediction Report

Report by- Shivangi Sharma

Problem Statement:

Please ensure that your final submission includes the following:

1. Solution file containing the predictions for the row_id in the test set (Format is given in sample_submission.csv)
2. A zipped file containing code & approach (Note that both code and approach document are mandatory for shortlisting)
a. Code: Clean code with comments on each part
b. Approach: Please share your approach to solve the problem (doc/ppt/pdf format). It should cover the following topics: i. A brief on the approach used to solve the problem.
ii. Which Data-preprocessing / Feature Engineering ideas really worked? How did you discover them?
iii. What does your final model look like? How did you reach it?

Introduction:

It doesn't get much better than customer churn prediction when it comes to meaningful commercial uses of machine learning. It's an issue with a lot of high-quality, fresh data to work with, it's reasonably simple to solve, and it may be a wonderful way to boost revenues.

The churn rate has a demonstrable impact, thus we need measures to lower it. Predicting churn is an effective technique to conduct preemptive marketing efforts for consumers who are likely to leave.
Customer turnover may be forecasted with the aid of machine learning thanks to big data. Machine learning and data analysis are practical tools for detecting and forecasting churn. You're also doing the following during churn prediction:
Identifying at-risk customers,
Identifying customer pain points,
Identifying strategy/methods to lower churn and increase customer retention.

i. A brief on the approach used to solve the problem.

About DataSet:

	ID	Age	Gender	Income	Balance	Vintage	Transaction_Status	Product_Holdings	Credit_Card	Credit_Category	Is_Churn
0	84e2fcc9	36	Female	5L - 10L	563266.44	4	0	1	0	Average	1
1	57fea15e	53	Female	Less than 5L	875572.11	2	1	1	1	Poor	0
2	8df34ef3	35	Female	More than 15L	701607.06	2	1	2	0	Poor	0
3	c5c0788b	43	Female	More than 15L	1393922.16	0	1	2	1	Poor	1
4	951d69c4	39	Female	More than 15L	893146.23	1	1	1	1	Good	1

ID: Customer ID unique for each customer
gender: Whether the customer is a male or a female
Age: Age of customer
Income: Income of the customer
Balance,Vintage,Transaction_Status,Production_Holdings,Credit_Card,Credit_Category, Is_Churn are the column of our dataset.

Defining problem and goal: The total scope of developing a machine learning-powered application to predict customer turnover follows a traditional ML project structure, which comprises the following steps:
Identifying the problem and the desired outcome: Understanding what insights you need from the analysis and forecast is critical. Gather needs, stakeholder pain points, and expectations to better understand the situation.

Preparing, exploring, and preprocessing data: To solve the problem and construct prediction models, raw historical data must be translated into a format appropriate for machine learning algorithms. This step can also help to enhance overall outcomes by improving data quality.

Modeling and testing: This includes creating and validating customer churn prediction models using a variety of machine learning methods.

Deployment and monitoring: This is the final step in using machine learning to anticipate churn rates. The most appropriate model is then pushed into production. It may be incorporated into current software or used as the foundation for a new application.

The pipeline used for this example consists of 8 steps:

- Step 1: Problem Definition
- Step 2: Data Collection
- Step 3: Exploratory Data Analysis (EDA)
- Step 4: Feature Engineering
- Step 5: Train/Test Split
- Step 6: Model Evaluation Metrics Definition
- Step 7: Model Selection, Training, Prediction and Assessment
- Step 8: Hyperparameter Tuning/Model Improvement

ii. Which Data-preprocessing / Feature Engineering ideas really worked? How did you discover them?

The following steps are defined to preprocess the features for machine readability and subsequent analysis based on the data kinds and values:

Columns removed:
Id column was removed as it has no contribution to our model.
Label encoding:
We used Label encoding to convert yes/no or male/female columns to convert into a numerical value. We changed female to 1 and male to 0 using replace function.
One hot Encoding: The following characteristics are categorical, but not ordinal (no ranking), and can have several values. A new variable is generated for each value, with a binary integer indicating whether the value occurred in data input or not (1 or 0).
Scaling from minimum to maximum The numerical characteristics' values are rescaled between 0 and 1. The min-max scaler is the most common scaling method. Standard scaler, which scales data around a mean of 0 and a standard deviation of 1, might be used for regularly distributed characteristics. We utilize a min-max scaler for all numerical characteristics to keep things simple.

iii. What does your final model look like? How did you reach it?

After completing the data preprocessing part we tried to predict using a linear regression model.
Using linear regression we get an accuracy of 0.7772209567198177 for the training data set.
Then we used As Decision Trees but using this algorithm we faced the problem of overfitting so we moved on to Random Forest Classifier.
Using Random Forest Classifier we get an accuracy of around 97% on training data and around 70 on test data.
We used n_estimators=500, max_depth=15 for our classifier.