Customer Churn Risk Prediction

Background

Your client is a major mobile telecommunication network provider in the US. They are experiencing issues related to customer churn or attrition i.e. customers cancelling their accounts and possibly switching to other competitor services.

Data

The data available includes customers' demographic profile, their plan features and usage history along with an indicator whether they actually churned or not. This is provided in the dataset below:

 Churn History Dataset.csv

Use this historical dataset to build/train the churn models. Then evaluate the prediction accuracy of the models on the test dataset below:

 Churn Test Dataset.csv

Objective:

The client is interested in understanding the leading indicators of churn and identifying potential churners ahead of time. This will enable them to take pre-emptive action such as offering better plans and discounts to potential churners and encouraging them to continue their service. The client has hired you to analyse their customer data and build models to estimate the churn probability for each active customer and generate useful insights.

Please implement the following modeling techniques and identify the "best" model for each technique:

I. Logistic Regression
II. Decision Trees (if possible)
III. Random Forests (if possible)

Here are some of the questions you need to think about and address:

I. Perform exploratory analysis and understand and present your findings. Which attributes seem to have a bearing on churn behavior? Any data transformations needed for modeling? Are there any outliers or extreme values that could skew your analysis? How do you want to deal with them?

II. How would you compare the performance of different models on the test dataset? What metrics and visualizations would you use to present your findings to the client?

III. If you had to pick one model/approach to recommend to the client, which one would you choose? Why? What would be your considerations? How would you justify your decision?

IV. What are the leading indicators of customer churn? Could you rank them in the order of importance?

V. How would you flag each customer in the test dataset as HIGH or LOW churn risk based on their predicted churn probability?