Project Proposal

Title of the project: Predicting Customer Churn for Telecom

Background and motivation / description of the problem:
Predicting *customer churn* (also known as customer attrition) is important because across all industries, businesses are not only concerned with attracting new customers but also with retaining current customers. If a business can better understand what factors may cause a person to be more likely to leave, then the business can adjust its marketing and corporate strategy in hopes of lowering their customer churn. The telecommunications industry in particular is especially concerned with decreasing customer churn since, in most countries, a customer can choose from several different companies and the process of switching telecom providers is fairly easy. The ability to accurately predict which customers are more likely to churn can lead to significant monetary benefits for a company, because reducing customer churn is a significantly less costly option than increasing the number of new customers.

Question you attempt to answer, functions you attempt to implement, etc. :

What type of customers are more likely to churn - based on our models. Which parameters are important which are not? Use functions to remove highly correlated features to each other, redundant features like customer ID, and function for tuning parameters if any for our models have them.

General direction of your solution. (for example, are you going to use classification? Are you going to use clustering? You can do multiple!!)

We plan to use classification as the data is labelled 'Yes' or 'No' (customer churns or not) and we want to classify new data with our model which is constructed by the training set. We can use classification algorithms like Neural Networks, Decision Tree, Logistic Regression, Random Forest. We try different algorithms, take the best one and tune it to get a better precision.

Describe the potential datasets to be used in experiments. How will you measure the performance of the proposed method? Are you going to measure the performance in terms of accuracy, efficiency, scalability? What's the reason?

Dataset:

It is a binary labelled dataset (Yes or No) for Churn column with other information about the customer like gender, age, married, using other services etc. We are going to measure performance by:

- ***Accuracy*** - Important but as We have few objects for yes (2/7th) and thus imbalanced classes, we guess 'No' everytime and get 70% accuracy. So we have to take other figures as well.

- **Weighted F1-Score:** Manage trade-off between that the Customer churned (Precision) and finding as many churned customers as possible (Recall). Especially important due to class imbalance in our dataset. So a model does well in F1 score if the positive predicted are actually positives (precision) and doesn't miss out on positives and predicts them negative (recall).F1-Score with higher weightage to precision to minimize cost when trying to retain churning customers by knowing exactly what type leave.

- **AUC-ROC**: Is used for binary-classification where it deals with situations with very skewed sample distribution, and do not want to overfit to a single class eg. No for Customer Churn.