DATA SCIENCE ASSIGNMENT

Customer Churn

Dataset: The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents.

Design choices: We have chosen two models which are going to evaluate and predict the dataset. The two models here are decision tree classifier and random forest classifier. One way to think of a Machine Learning classification algorithm is that it is built to make decisions. You usually say the model predicts the class of the new, never-seen-before input but, behind the scenes, the algorithm has to decide which class to assign.

Decision tree classifier- is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. They are simplest algorithm. They are generally faster than other neural networks and can handle high dimensional data.

Random forest classifier – Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.

Performance evaluation of the model

The performance evaluation for both the models is done through model testing and found following evaluation.

Decision tree classifier – the accuracy of test data in this model is 79% which is above the required objective.

Random forest classifier- the accuracy of test data in this model is 78%.

Discussion of future work – we can actually try many other classifier models and select the best model depending on their accuracy percentage. Also we can thin out data and perform certain specific analytical actions. Also we can further build prediction model which takes out live dataset and predict the accuracy continuously.

Tasks achieved in my assignment

- Classification labels/Targets Variables: Churn — Whether the customer churned or not (Yes or No), Tenure — Number of months the customer has stayed with the company, Reasons - Reasons behind the customer churn (Multiple Reasons possible), NA if Churn = No
- Hypothesis Building should be concise
- Deploy max 6 most important features after applying most suitable feature reduction techniques
- Accuracy of the model on the test data set should be > 70%
- Add at least one method for Hyperparameter tuning.
- Add at least one method for model validation other than Train/Test Split

Source code

I have given a collab link below and also shared my code in file.

https://colab.research.google.com/drive/10wUaq3Fs5rWRYXa4bbw_9O2M6Y8O4Yu7?usp=sharing