ML Engineer HW Assignment

Introduction

At Carta, we are highly focused on the growth of our business. One of the most important parameters for net revenue growth of SaaS businesses is churn prevention.

In this exercise, we will explore a publicly-available dataset to build a predictive model for detecting customers that are at risk of churning, i.e., classifying whether an existing customer is likely to stop using and paying for a product. While the assignment dataset is not for a SaaS company, many techniques and concepts are transferable.

You can use any set of tools that you like (R, Python, etc.), but we intend for you to spend about two and a half hours on the assignment.

Dataset

The attached dataset (`churn_dataset_train.csv`) contains information about customers of a telecom company with a binary label indicating whether an individual has churned or not. The `churn_dataset_test.csv` file contains unlabeled data and will be used to generate out-of-sample predictions.

The following **features** (explanatory variables) are available:

- ***state*** (customer's state of residence)
- ***account_length*** (no. of months customer has been with the provider)
- ***area_code*** (3 digit area code)
- ***international_plan*** (binary indicator of whether a customer has international plan)
- ***voice_mail_plan*** (whether a customer has voice mail plan)
- ***number_vmail_messages*** (no. of voice-mail messages)
- ***total_day_minutes*** (total minutes of day calls)
- ***total_day_calls*** (total number of day calls)
- ***total_day_charge*** (total charge of day calls)
- ***total_eve_minutes*** (total minutes of evening calls)
- ***total_eve_calls*** (total number of evening calls)
- ***total_eve_charge*** (total charge of evening calls)
- ***total_night_minutes*** (total minutes of night calls)
- ***total_night_calls*** (total number of night calls)
- ***total_night_charge*** (total charge of night calls)

- ***total_intl_minutes*** (total minutes of international calls)
- ***total_intl_calls*** (total number of international calls)
- ***total_intl_charge*** (total charge of international calls)
- ***number_customer_service_calls*** (no. of calls to customer service)

The **variable to be predicted** is the column ***churn***. It can assume two values: “yes” and “no”.

Evaluation

Your submission will be evaluated on the following four criteria:

1. Is your code **readable**?
2. Are the **data science techniques** used appropriate for solving the problem and evaluating generalized performance?
3. Is there a **clear narrative** around the findings and an explanation of methodology choices.
4. How well does your model generate **predictions** for out-of-sample data?

Questions

In addition to your solution, please submit a brief write-up answering the following three questions:

1. Did you preprocess any of the features? If so, why? If not, why not?
2. Which features are the most relevant for predicting the output? How did you measure feature importance?
3. What metrics did you use to measure the performance of your model? How did you determine how well your model generalizes?

Submission

Submit your work as a zip file (no GitHub links, please!) named LASTNAME-CARTA.zip.

Your submission should include the following files:

1. `predictions_churn_dataset_test.csv`: this file should be generated by filling the column churn in the file `churn_dataset_test.csv` with your predictions ("yes" or "no") for each unlabeled instance. Keep all the columns, including the index one, so that we can identify each sample when checking your results.
2. `writeup.pdf`: a document containing your answers to the above questions (this could be a Word or Google Doc saved as a PDF).
3. `code.[file extension]`: your code, in whatever format is most appropriate. Your code should include all the steps for building your model, from loading the training dataset to generating predictions.

Please do **not** include the `churn_dataset_train.csv` file in your submission.