**Banking Customer Churn Project**

You work as a data scientist at a US bank. The bank offers a special savings account intended for middle income professionals. Interest is accrued in the savings account on a monthly basis and is only accrued if there is a net positive change after deposits and withdrawals in that month. A small fee is charged with customers withdraw from the account. Customers are therefore incentivized to continually deposit into the account to take advantage of the interest.

The bank uses these deposits to fund activities and investments. It is in the bank’s interest to keep the savings portfolio healthy as a larger savings portfolio gives the bank a larger liquidity buffer. It has been identified that customer churn is one of the main sources of outflow of funds from the bank.

Your manager wants to identify customers who are likely to churn and see whether they can be incentivized to keep their accounts open. Your task is to develop a model to predict the likelihood of customer churn and convince others that your model and analysis are trustworthy.

**Dataset Description**

**Date Range**

*Start date* : 2007-01-31

*End date* : 2020-05-31

You want to identify customers who close their accounts in the following month.

**Files**

You have been provided two files by the data team:

|  |  |
| --- | --- |
| **File** | **Description** |
| transactions\_tm1\_e.csv | File containing saving account transactions for select customers. |
| customers\_tm1\_e.csv | File containing customer data. |

**Task 1: Exploratory Data Analysis**

Before fitting our model, you will need to make sure our data is fit for purpose. There may also be some data pre-processing required before the data is in a form suitable as input for a supervised learning model.

**Data Integrity Checking**

The first step in our exploratory data analysis and ensure that the data is fit for purpose.

Some things to consider include:

* What visualizations could be useful for an initial sanity check?
* Does the data look like you expect?
* Are there any outliers that could skew your results?

**Data Pre-Processing**

Consider how you can construct your matrix of features and your label vector.

**Exploratory Analysis**

Here you will explore the data and get a feel for what the data looks like.

Some things to consider include:

* Can you use any visualizations to better understand the data?
* Are there any overall trends present?
* What does the customer demographic look like?
* Has the portfolio grown or shrunk over time?
* Is there a relationship between the data and external data?
* Does the churn rate change over time?

**Relationship to Macroeconomic Data**

External data you can consider include macroeconomic data from the Federal Reserve. The data can be obtained here: <https://fred.stlouisfed.org/>. You can also consider US population distributions to see whether this lines up with the data.

Some macroeconomic variables to consider include:

* University of Michigan Consumer Sentiment Index
* GDP
* Inflation
* Unemployment Rate
* Interest Rates
* Income

Think about what relationships you should expect, and whether the data is consistent with your expectations.

**Task 2: Modelling**

The task here is to build a model to predict customer churn in the following month. Think about what features are predictive of churn and experiment to determine whether they help with your predictions.

Use your model to make predictions of whether active customers in the current month are likely to churn. Your predictions will be used to validate your model.

Note: Use of Azure AutoML/Cognitive Services is not allowed for this project.

**Task 3: Model Interpretation**

You are required to be able to explain and interpret your model. For model interpretation, you can consider techniques such as partial dependence plots and SHAP.

**Task 4: Report**

Your team is required to submit a report outline your analysis and modelling. You will need to demonstrate understanding of the data and describe how you achieved your final model. You should also attempt to interpret what your model is doing.

It is recommended that you invest time thinking about a clear structure for the report. As a guideline, the report should be no more than 20 pages in length, but please keep it as concise as possible.

**Submission**

For this project, you will need to submit both a report and a submissions file. Your submissions file will contain the propensity (probability) to churn for all active accounts in the dataset.

For your submission, please include the following and use the suggested file names:

- Report submission in pdf format (churn\_report\_<pod\_name>.pdf).

- Submission csv file (submission\_<pod\_name>.csv). Make sure you have the correct column names.

- Presentation in pdf format (churn\_presentation\_<pod\_name>.pdf).

Please submit to the usual submissions email ([mle<xx>submissions@kubrickgroup.com](mailto:mle%3cxx%3esubmissions@kubrickgroup.com)) using the title "MLE<xx> Churn Project - Pod <pod\_name> Submission".

In your email, please also include a link to your git repo. Let us know if there are any questions.

**Grading**

Your final grade will be based on a combination of the quality of your report, your predictions, engagement, and your presentation.

**Hints**

* You may need to aggregate by month. For consistency, you can use the pandas timestamp datatype for dates and use the month end convention.
* Make sure you try to understand the data provided to you.
* Always perform sanity checks on your data before moving to the modelling stage.
* Think about what model evaluation metrics will be suitable.
* Make sure to engage and ask questions. There may be information you can obtain outside of this specification.