FIT5149 TP6 2021 Assessment 1: Customer Churn Prediction

Marks	20% of all marks for the unit
Due Date	23:55 Sunday 14 November 2021
Extension	An extension could be granted for circumstances. A special consideration application form must be submitted. Please refer to the university webpage on special consideration.
Lateness	For all assessment submissions handed in after the official due date, and without an agreed extension, a **10% penalty** applies to the student's mark for each day after the due date (including weekends, and public holidays) for up to 5 days. **Assessment items handed in after 5 days will not be considered.**
Authorship	This assignment is an individual assignment and the final submission must be identifiable as your own work. Breaches of this requirement will result in an assignment not being accepted for assessment and many result in disciplinary action.
Submission	You are required to submit two files, one is either a Jupyter notebook or a R Markdown file, another is the PDF file generated by them. The two files must be submitted via Moodle. Students are required to accept the terms and conditions in the Moodle submission page. A draft submission won't be marked.
Programming language and tools	R in Jupyter Notebook or R Markdown

1. Introduction

Insurance companies around the world operate in a very competitive environment. With various aspects of data collected from millions of customers, it is very hard to analyze and understand the reason for a customer's decision to switch to a different insurance provider. For an industry where customer acquisition and retention are equally important, and the former being a more expensive process, insurance companies rely on data to understand customer behavior to prevent retention.

Thus knowing whether a customer is possibly going to switch beforehand gives Insurance companies an opportunity to come up with strategies to prevent it from actually happening. In this assessment, you are given 16 distinguishing factors that can help in understanding the customer churn. Your objective as a data scientist is to build a Machine Learning model that can predict whether the insurance company will lose a customer or not using these factors. You are provided with 16 anonymized factors (feature_0 to feature 15) that influence the churn of customers in the insurance industry. The datasets are splitted as follows:

- Train.csv – 27126 observations.
    - Predictors: 16 anonymized factors (feature_0 to feature 15)
    - Target variable: Label ('1' indicates churn)
- Test.csv – 6782 observations.

In this assessment, the aim is to build statistical learning models that can predict the customer churn. Specifically, the problem you are going to solve is:

- Can you accurately predict the customer churn for the insurance company given the collected data?
- Can you well explain your prediction and the associated findings? For example, identify the key factors that are strongly associated with the response variable, i.e., the Label.

2. Task description

In this assessment, you will focus on the following two tasks.

2.1 Prediction task

For the prediction task, the underlying problem is to predict the customer churn using the collected attributes. The provided dataset is well organised and cleaned. It is important that you understand the importance of each attribute in the problem.

To measure the performance of your model(s), you should fit the model to the training dataset, perform the prediction on the test dataset and finally compute some performance metrics of your choice. At least you should use two performance metrics to score the models. **We expect you to justify these metrics according to the data characteristics.** To successfully finish the task, you should:

1. Perform a comprehensive Exploratory Data Analysis (EDA) on the data to identify the features' characteristics and relevance. At least uni- and bi-variate feature analysis are expected;
2. Describe and justify the choice of your models according to your EDA;

3. Develop two types of models (two different modelling families (e.g., KNN and Logistic Regression)) on the train set using train_test_split or K-fold cross validation strategy. **Models with different feature sets or hyper-parameters are not considered as two different families** (e.g., “two KNN models with k=2 and 3” or “two Logistic Regression models with different features” are not accepted);
4. Evaluate your models using your selected performance measures. Then, analyze and compare your models’ performance according to their scores on the test set and models characteristics (such as train time, response time, etc.) and select one of two developed models as your final model.

2.2. Description task

The purpose of the description task is to give a clear insight to why customers churn. Descriptions can be based on variable correlation analysis, regression equations, linguistic descriptions, or any other form. The descriptions and accompanying interpretation must be comprehensible, useful and actionable for a marketing professional with no prior knowledge of computational learning technology. To finish this task, yow need to:

1. Perform model inference on the selected model from previous step, identify and discuss the key factors (variable importance) of the selected model;
2. Make your suggestions in terms of marketing.

3. Files to be submitted

There are two files required to be submitted, which are

- The R implementation of the two tasks in one file.
  - The file must be either a Jupyter notebook or an R Markdown file. Besides the R code, all the discussions must also be included in the file.
  - The name of the file must be in one of the following formats: “**XXXXXXXX FIT5149 Ass1.ipynb**” or “**XXXXXXXX FIT5149 Ass1.Rmd**”. You should replace “XXXXXXXX” with your student ID.
- A PDF file generated by the Jupyter notebook or R Markdown.

Please refer to the Assessment 1’s Moodle page for how to submit the two files.

4. Additional learning resources

This task is based on this Kaggle competition challenge (the provided dataset is modified). You can use its Kernels to get some ideas but remember you need to cite any external sources that you use in your notebook

Warning: Monash University takes academic misconduct very seriously. You can learn from the above materials and understand the principle of how the analysis was done. However, you must finish this assessment with your own work.