**Data Science Project Protocol**

*Author(s):*

# Introduction

Here you have to give some known facts about the field you will work on.

Try to focus on the problems that are most common and then state the goals of the project.

* You can try to answer to the following question:
* Which questions do we want to answer ?
* What is known about the problem?
* How we define the outcome(s)?
* What is known to influence the outcome?
* Does we have any possible new knowledge that has not been in use before?

This part must be between half to one and half page.

an emphasis on practical but principled novel models.

[KKBOX](https://www.kkbox.com/) is Asia’s leading music streaming service, holding the world’s most comprehensive Asia-Pop music library with over 30 million tracks. They offer a generous, unlimited version of their service to millions of people, supported by advertising and paid subscriptions. This delicate model is dependent on accurately predicting churn of their paid users. For a subscription business, accurately predicting churn is critical to long-term success. Even slight variations in churn can drastically affect profits.

Customer churn occurs when customers or subscribers stop doing business with a company or service, also known as customer attrition. It is also referred as loss of clients or customers. One industry in which churn rates are particularly useful is the telecommunications industry, because most customers have multiple options from which to choose within a geographic location.

Our purpose is to build an algorithm that will predict churn for KKBOX’s paid subscriptions after their subscription expires.

The churn/renewal definition can be tricky due to KKBox's subscription model. Since the majority of KKBox's subscription length is 30 days, a lot of users re-subscribe every month manually.

# Methodology (Project design)

## Data

Here you have to describe how do you plan to manipulate the data. For this you have to answer to the following questions:

* Which data will be used?
  + Describe data sources
  + Describe possible external data sources that may enrich our data
  + Data for external validation?
* On which time frames periods will your project will be based on?
  + Time-frame for training
  + Time-frame for test?
* How do you define your subjects?
  + Inclusion criteria?
  + Exclusion criteria?
* Which would be your outcome variable?
* Are there confounder variables that may affect the outcome?
* Is there a possible source of bias in our data?
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How you will deal with outliers?
* How you will deal with missing values
* Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* What techniques will you apply to model your outcome?
  + Unsupervised
  + Regression
  + Classification
* Will you use cross-validation and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently the model will be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

Here you will present the main results of all the process. We will describe:

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new features.

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss also the limitations of the model, when it can be used and when not.