CE802-7-AU-CO Machine Learning

Dr. Luca Citi

Assignment 1: Pilot-Study Proposal

Student ID: 2003838 – PEREZ62206

January 2021

(Note: This document has 979 words in total. However, 16 of them are for the cover page, 168 of them are for the references, and 45 of them are for this note.

The rest **750** **words** are **designated** to the **narrative** of this pilot-study proposal)

**Proposed approach**

The machine learning procedure proposed in this pilot-study is summarized in the following sections:

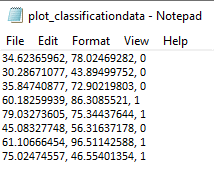
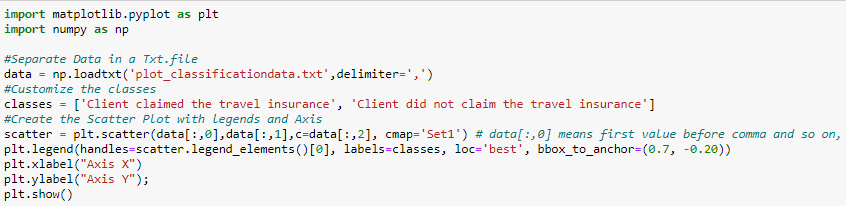
1. **The type of predictive task that must be performed.**

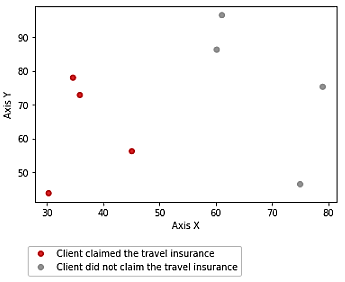
In this project, the manager of a travel-insurance company is facing with a problem that requires an automated decision: is a customer likely to claim? (yes/no).

In other words, the manager has a classification problem, where there are 2 categories:

(client will claim/client won’t claim) (See figure 1.0)

Figure 1.0:





**Source**: Own elaboration in Python 3 to illustrate the classification problem. (Data is fictitious)

In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes (Alpaydin 2014), and it is, perhaps, the most important form of prediction, which the goal is to forecast whether a record is 0 or 1: (Bruce and Bruce 2017).

As a result of this, a machine learning algorithm for classification**[[1]](#footnote-1)** will be suggested to successfully solve this problem.

1. **The Machine Learning procedure that could be performed.**

As mentioned above, the manager has a binary-categorical classification problem, where the output has two categories:

1. Customer will claim its travel assurance (*y*) given certain features (*x*)
2. Customer won’t claim its travel assurance (*y*) given certain features (*x*).

In machine learning, logistic regression is a widely recognized classification method that is used to predict the probability of a binary-categorical dependent variable **[[2]](#footnote-2)**. (Pesantez-Narvaez, Guillen, and Alcañiz 2019). In other words, as Li (2017) states, a logistic regression model will “identify relationships between our target feature (will claim/won’t claim), and our remaining features to apply probabilistic calculations for determining which class the customer should belong to.

Since the manager wants to make a binary-categorical prediction, a logistic regression model will be proposed to forecast whether a customer will claim its travel assurance or not (y).

1. **Examples of** **possible informative features.**

From now on, the manager must take a data sample, which should contains any of the following features described in Table 1.0

Table 1.0









**Source**: Own elaboration in Excel to illustrate the possible features that the travel insurance company should provide

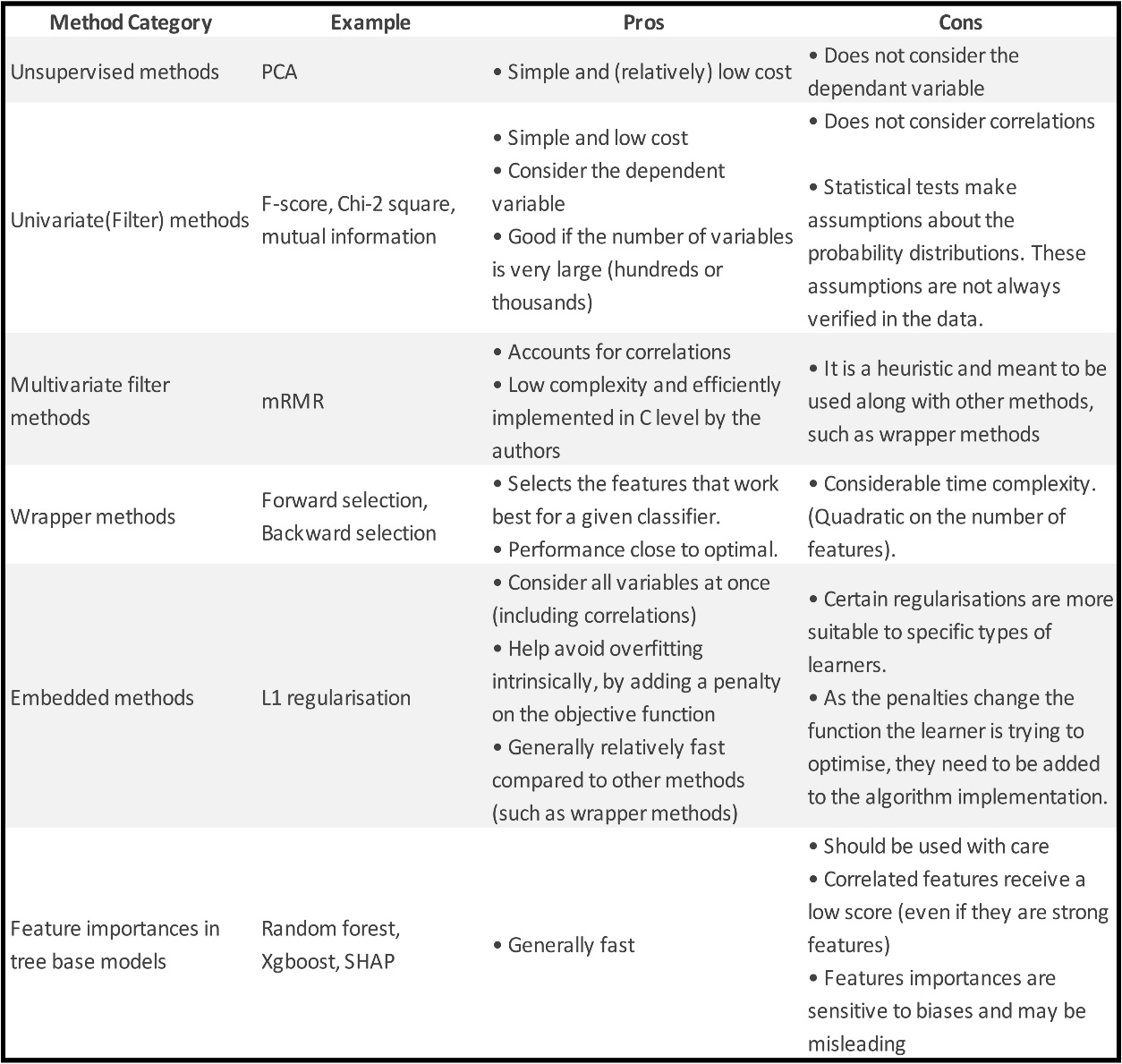
1. **The performance evaluation before deployment that could be performed.**

Once a dataset (containing possible informative features) has been provided by the manager, a logistic regression model will be developed and then deployed. However, before to do that, it is very important to evaluate the performance of the model built.

Before to train and test the logistic regression, a feature selection technique can be applied to reduce the number of classes in the model and optimize its performance. As Azevedo (2019) states: “having irrelevant features in your dataset can decrease the accuracy of the model”.

Since logistic regression is a classification method, it is possible to choose any of the most common types of feature selection techniques for classification problems (see table 2.0).

Table 2.0



**Source**: Adapted from (Azevedo 2019).

However, there are three ways in logistic regression to rank features: (Data Detective 2019):

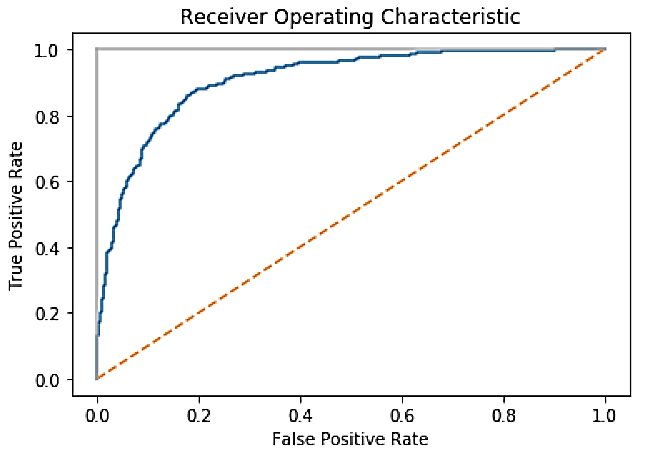
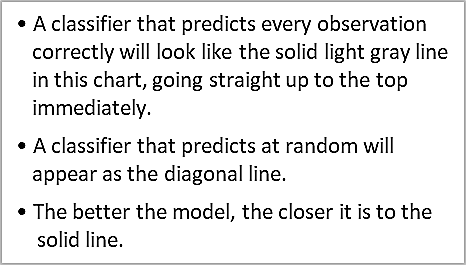
1. Recursive Feature Elimination (RFE) **[[3]](#footnote-3)**
2. Coefficient values
3. Sci-kit tool: SelectFromModels (SFM)

In this case, the RFE technique will be used because “it is easy to configure and effective at selecting those features that are more relevant in predicting the target variable” (Brownlee 2020).

Finally, after training and testing the logistic regression, the essential model evaluation technique for binary classification **[[4]](#footnote-4)** should be applied to find the effectiveness of the algorithm built.

In this case, the Receiving Operating Characteristic (ROC) Curve method will be used to evaluate how good the logistic regression is in predicting the outcome of new observations. As (Albon 2018) states: “by plotting the ROC curve, we can see how the model performs because it compares the presence of true positives and false positives at every probability threshold”. (see figure 2.0).

Figure 2.0



**Source**: Adapted from (Albon 2018)

**References**

Albon, Chris. 2018. *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*. Sebastopol, CA: O’Reilly Media.

Alpaydin, Ethem. 2014. *Introduction to Machine Learning*. Third edit. Cambridge, Massachusetts: The MIT Press.

Asiri, Sidath. 2018. “Machine Learning Classifiers. What Is Classification?” *Towards Data Science*. Retrieved December 25, 2020 (https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623).

Azevedo, Gabriel. 2019. “Feature Selection Techniques for Classification and Python Tips for Their Application.” *Towards Data Science*. Retrieved December 28, 2020 (https://towardsdatascience.com/feature-selection-techniques-for-classification-and-python-tips-for-their-application-10c0ddd7918b).

Brownlee, Jason. 2020. “Recursive Feature Elimination (RFE) for Feature Selection in Python.” Retrieved December 28, 2020 (https://machinelearningmastery.com/rfe-feature-selection-in-python/).

Bruce, Peter, and Andrew, Bruce. 2017. *Practical Statistics for Data Scientists*. Sebastopol, CA: O’Reilly Media.

Data Detective. 2019. “A Look into Feature Importance in Logistic Regression Models.” *Towards Data Science*. Retrieved December 28, 2020 (https://towardsdatascience.com/a-look-into-feature-importance-in-logistic-regression-models-a4aa970f9b0f).

Li, Susan. 2017. “Building A Logistic Regression in Python, Step by Step.” *Towards Data Science*. Retrieved December 25, 2020 (https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8).

Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. “Predicting Motor Insurance Claims Using Telematics Data—XGboost versus Logistic Regression.” *Risks* 7(2). doi: 10.3390/risks7020070.

1. The most common machine learning algorithms for classification are:

   Logistic Regression, K-Nearest Neighbours, Support Vector Machines, Kernel SVM, Naïve Bayes, Decision Tree

   Classification & Random Forest Classification (Asiri 2018)**.** [↑](#footnote-ref-1)
2. Binary-categorical data isa data that can take on only a specific set of 2 values (will claim/won’t claim) representing a set of possible categories (Customer\_Insurance\_Claim)(Bruce and Bruce 2017). [↑](#footnote-ref-2)
3. RFE is a “feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached” (Data Detective 2019) [↑](#footnote-ref-3)
4. According to Albon (2018), the essential metrics and methods used for assessing the performance of predictive binary classification models, includes: **Average classification accuracy, Confusion Matrix, Precision, Recall, Specificity and ROC Curve.** [↑](#footnote-ref-4)