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:

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

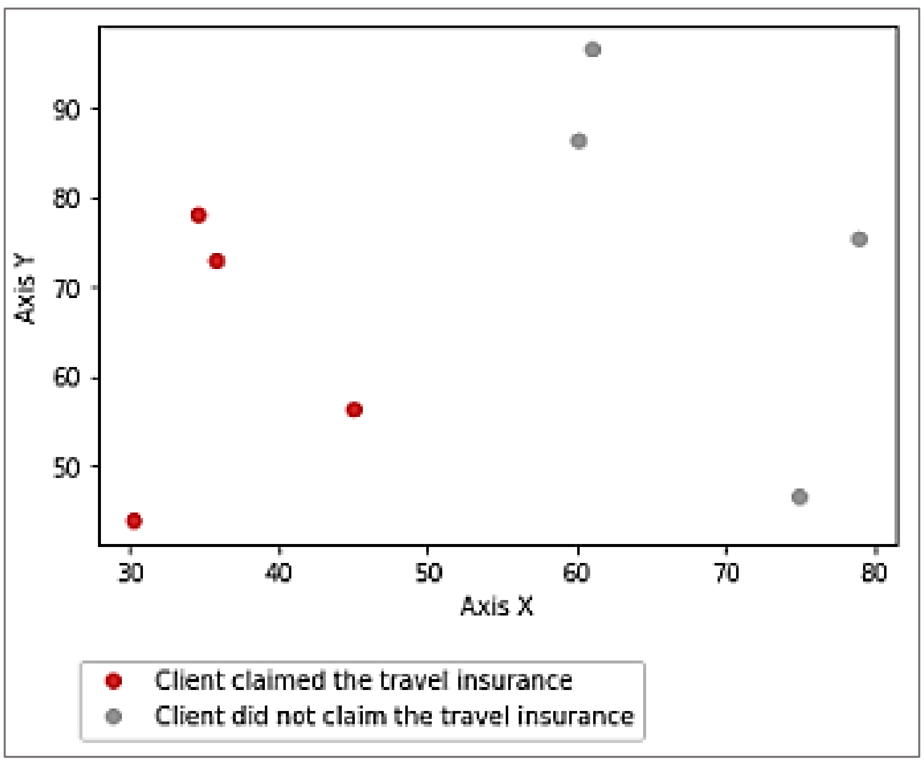
In this project, the manager of a travel-insurance company is facing with a problem that requires an automated decision: is a customer likely to claim? (yes/no).

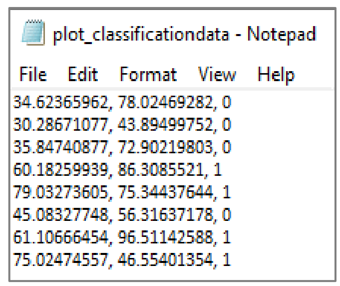
In other words, the manager has a classification problem, where there are 2 categories: (client will claim/client won't claim) (See figure 1.0)

Figure 1.0:

```python
import matplotlib.pyplot as plt
import numpy as np

#Separate Data in a Txt.file
data = np.loadtxt('plot_classificationdata.txt',delimiter=',')
#Customize the classes
classes = ['Client claimed the travel insurance', 'Client did not claim the travel insurance']
#Create the Scatter Plot with legends and Axis
scatter = plt.scatter(data[:,0],data[:,1],c=data[:,2], cmap='Set1') # data[:,0] means first value before comma and so on,
plt.legend(handles=scatter.legend_elements()[0], labels=classes, loc='best', bbox_to_anchor=(0.7, -0.20))
plt.xlabel("Axis X")
plt.ylabel("Axis Y");
plt.show()
```





plot_classificationdata - Notepad

File Edit Format View Help

34.62365962, 78.02469282, 0
30.28671077, 43.89499752, 0
35.84740877, 72.90219803, 0
60.18259939, 86.3085521, 1
79.03273605, 75.34437644, 1
45.08327748, 56.31637178, 0
61.10666454, 96.51142588, 1
75.02474557, 46.55401354, 1

**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] will be suggested to successfully solve this problem.

## B. 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]. (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).

## C. 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

[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).

[2] Binary-categorical data is a 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).

Table 1.0

| Information about the Insured | | |
|---|---|---|
| **Feature** | **Type of Feature** | **Brief description** |
| Age | Continuous | Age of the policyholder (client).<br>* For this feature, it should be considered a **Binning** or **Discretization** technique. |
| Nationality | Categorical | Nationality of the policyholder (client)<br>* For example: Mexican, British, Chinese, Indian, American, etc. |
| Gender | Categorical | Gender of the policyholder (client)<br>* For example: Female, Male |
| Dependents | Categorical | Whether customer has dependents or not.<br>* For example: Yes, Not |

| Information about the Insurance | | |
|---|---|---|
| **Feature** | **Type of Feature** | **Brief description** |
| Product | Categorical | The type of travel-insurance bought by the policyholder<br>* For example: Basic Insurance, Baggage Insurance, All-Included Insurance, etc. |
| Payment | Categorical | The policyholder's payment method.<br>For example: E-Check, Mailed Check, Bank Transfer, Credit Card, etc |
| Tenure | Numerical | The Number of months the policyholder has been with the Travel Insurance Company<br>* For example: 12, 8, 9,48,1,0 |
| contract | Categorical | The Term of the policyholder's contract<br>* For example: Monthly, 1-Year, 2-Year |

| Information about the Flight | | |
|---|---|---|
| **Feature** | **Type of Feature** | **Brief description** |
| Counts_Baggage | Numerical | The number of baggage carried in a flight per policyholder<br>* For example: 1,2,3 |
| Origin | Categorical | The city where the policyholder starts the journey<br>* For example: Mexico City, London, Paris, Rome, Madrid |
| Destination | Categorical | The city where the policyholder ends the journey<br>* For example: Mexico City, London, Paris, Rome, Madrid |
| Purpose | Categorical | The reasons for traveling provided by the policyholder<br>* For example: Business, Tourism, Studies, Pleasure, etc. |
| Airline | Categorical | The airline taken by the policyholder<br>* For example: British Airways, Qatar Airways, EasyJet, Ryanair, etc. |
| Duration | Categorical | The length of time the policyholder will spend flying expressed in 1 categorie<br>* For example: Short Haul [flight < 3 hrs), Medium Haul [3hrs < flight < 6hrs],<br>Long Haul [6hrs < flight < 12hrs] & Ultra Long Haul [flight > 12hrs] |

| Target Feature | | |
|---|---|---|
| **Feature** | **Type of Feature** | **Brief description** |
| Insurance_Claim | Categorical | Whether the customer will claim the travel-insurance<br>* For example: Yes, Not |

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

**D. 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

| Method Category | Example | Pros | Cons |
|---|---|---|---|
| Unsupervised methods | PCA | • Simple and (relatively) low cost | • Does not consider the dependant variable |
| Univariate(Filter) methods | F-score, Chi-2 square, mutual information | • Simple and low cost<br>• Consider the dependent variable<br>• Good if the number of variables is very large (hundreds or thousands) | • Does not consider correlations<br><br>• Statistical tests make assumptions about the probability distributions. These assumptions are not always verified in the data. |
| Multivariate filter methods | mRMR | • Accounts for correlations<br>• Low complexity and efficiently implemented in C level by the authors | • It is a heuristic and meant to be used along with other methods, such as wrapper methods |
| Wrapper methods | Forward selection, Backward selection | • Selects the features that work best for a given classifier.<br>• Performance close to optimal. | • Considerable time complexity. (Quadratic on the number of features). |
| Embedded methods | L1 regularisation | • Consider all variables at once (including correlations)<br>• Help avoid overfitting intrinsically, by adding a penalty on the objective function<br>• Generally relatively fast compared to other methods (such as wrapper methods) | • Certain regularisations are more suitable to specific types of learners.<br>• As the penalties change the function the learner is trying to optimise, they need to be added to the algorithm implementation. |
| Feature importances in tree base models | Random forest, Xgboost, SHAP | • Generally fast | • Should be used with care<br>• Correlated features receive a low score (even if they are strong features)<br>• Features importances are sensitive to biases and may be misleading |

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

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

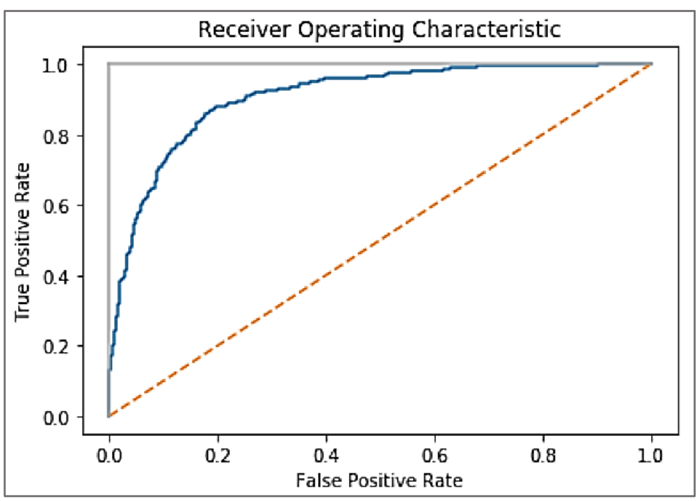
1. Recursive Feature Elimination (RFE) [3]
2. Coefficient values
3. Sci-kit tool: SelectFromModels (SFM)

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] 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




- A classifier that predicts every observation correctly will look like the solid light gray line in this chart, going straight up to the top immediately.
- A classifier that predicts at random will appear as the diagonal line.
- The better the model, the closer it is to the solid line.

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

---

[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)

[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.

## 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.