Task

As part of the exercise, you play the role of a Data Scientist working on a Proof of Concept engagement. Your (imagined) customer has sent you a sample dataset and tasks you to use ML to analyse the data and come up with a viable solution.

Dataset: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#

Dataset

The dataset consist of a xls file containing 30000 rows of information related to payments of credit card users from an important bank in Taiwan. The data contained the demographic information of clients, the bills and payments made by clients from April 2005 to September 2005, as well as the repayment status fields which records whether the payment was made on time or delayed and for how long. The dataset contains 24 Attributes in total.

Approach

The Dataset presented the opportunity to develop a Machine Learning model to predict the number of credit card clients that will likely be in default in the next month. The approach taken was to focus on ensemble methods, due to their flexibility and robustness, taking into account the medium-low size of the dataset, as well as their resilience to overfitting.

The first step after the EDA was to create a base model using Random Forest, and then compare it using XGBoost. Later there were added additional features to assess the changes from the base model.

Exploratory Data Analysis

The description of the data highlighted some errors in it that were not described in the Attribute information, these errors could have arisen from bugs or missing information during the data collection process.

First, the PAY_x status columns contained 2 more values (-2 and 0) which were not described, and due to the sensibility and limited scope of the data, it was not possible to make assumptions regarding the values.

Second, there were 317 rows of the data which contained BILL_AMTx values equal to 0, and for which the ‘default payment next month’ column was set to 1, indicating that the client was in default, however, for a client to be in default, the client must owe money to the bank. In addition, most of the attributes indicating the PAY_x

status contained values of -2, which its description was unknown. Therefore, those rows were dropped from the dataset.

Third, the data only overlaps by 5 months, as the bill from september, was paid in october, and the payment in April was made for the march bill, therefore, the only matching information corresponds from April to August in the BILL_AMTx columns, and from May to September in the PAY_AMTx columns.

According to the dataset, 39.63% of the customers were male and 60.37% were Female, and 22.12% of the customers were already on default.

From the histograms of the BILL_AMTx and PAY_AMTx columns, it was brought to attention that the information is skewed to the right and presents outliers above 500000 NT dollars and with some negative values for the BILL_AMTx columns, due to overpayments from customers which will create negative balances. From the comparison of the histograms, it can be deducted that higher the amount due to pay, the less payments are made.

From the Correlation heatmap, it was noticeable that the BILL_AMTx and PAY_AMTx, as well as the PAY_x and BILL_AMTx were positive correlated; which indicates that when one column goes up, the other one will go up, as it will be expected.

Modelling

The dataset was split between train and test datasets, with a quarter of the dataset being reserved for testing purposes. The first algorithm to be applied was Random Forest, that produced an accuracy of 81.62%, subsequent models using XGBoost showed a timid 0.09% and 1.38% of improvement after the addition of features.

In comparison with Random Forest, XGBoost presents less false positives, where the model predicted a No Default (0), but the actual outcome was Default (1). The False negatives are less worrisome for the bank, as customers that were in risk of going into default made a payment.

Implementation

For a successful implementation of the predictions generated by the Machine Learning models, the creation of a web application tool hosted in cloud services, could be useful to analysts aiming to target potential customers with high risks and plan ahead to negotiate a payment plan with the customers to avoid the credit card going into default. It could be designed to support the decision making process regarding the issue of new credit cards to new and existing customers by obtaining the relevant information and predictions through an API.

Furthermore, it can be offered to customers that wish to have a better control over their finances by providing them with a tool to make them self conscious about their spending and debt situation, and offer them re-payment products tailored to their needs, in a matter of minutes and few clicks.

Conclusions

According to the findings, the conclusions are summarised below:

- The economic impact of default in credit cards could mean millions in losses for banks and financial companies. The use of Machine Learning and Deep Learning are helping to tackle the problem and to minimise the risks associated with lending money. However, this tools are only as good as the data that they are ingested with, hence the quality and accuracy of the data collected needs to be prioritised and curated carefully.
- The base models provide a useful tool to predict and detect customers with risk of default in payments with an accuracy of approximately 82%. However the models can be subject to bias and other errors that could affect the integrity of the predictions, due to the nature of the data itself.
- In order to achieve a significant increase in accurate predictions, it is necessary to conduct further research and experimentation with the data, as well as experiment with different topologies and algorithms, including neural networks and deep learning, in order to assess the benefits and downsides of implementing them at production level.

Limitations

- These models are only experimental models, to create more advance models it will be required access to bigger databases and datasets from different aggregators, as well as further research and development to find an approach that will satisfy the requirements and needs of the client with the highest industry standards.

A copy of the code and repository can be found at:
https://github.com/jmiguel99/inawisdom.git