II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Anti-fraud system for detecting non-compliances in corporate spending

Luiz Gustavo Ribeiro[1], Fabricio Simeoni de Sousa[2]
Instituto de Ciências Matemáticas e de Computação - ICMC/USP, Campus de São Carlos

1 Introduction

Corporate expenses management is a vital part of several processes that guarantee the company's financial health. The bigger the company is, the harder this expenses management becomes. The use of software tools can facilitate a more in-depth analysis of the company's cash flow, such as expenses related to corporate trips and expense reimbursement.

Reimbursements are a special case consisting of the devolution of properly proven expenses by an employee. These expenses are generally made in external client meetings and business trips that cover food, lodging, plane or other transportation tickets, gas, and job material. A well-defined reimbursement policy is a key to ensuring transparency in the company's financial workflow and can help prevent non-compliance and fraud. In this case, non-compliance and fraud can be understood as any omission or intentional act promoted to harm and deceive the company.

Recent advances in artificial intelligence allow performing detection of non-compliant corporate expenses quickly and effectively, reducing the time needed for audits and increasing the reliability of the information. In this paper, we explore different mechanisms and apply different artificial intelligence algorithms to detect non-compliant corporate expenses to detect and identify cases of reimbursement fraud indirectly: We explore the dataset and describe the different operations to clean the data. We analyze the role of features in feature evaluation and selection. We apply several machine learning models for the prediction part of the work.

2 Material and Methods

2.1 Datasets

The database was constructed through several meetings with the owners of the system to understand which tables were pertinent. With the joining of these databases, a final database had

[1]lgribeiro@usp.br
[2]fsimeoni@icmc.usp.br

more than 600 thousand records, and with exploratory analysis, there was a great reduction to 230 thousand records. An expert created new attributes for this final database.

2.2 Selection of Attributes by Mutual Information and Feature Importance

Information theory was proposed in the 1940s by Claude E. Shannon [1], to study the transmission, storage, and quantification of information. It used the concepts of entropy and mutual information to identify and select variables that have the most information shared with the outcome of whether the register is a fraud.

The entropy measures the homogeneity of a given set. Depending on the probability of an event, the gain of information is characterized by the event's uncertainty since known events do not add anything. Entropy (H) is defined by the equation 1:

$$H(X) = -\sum_{x} p(x) \log_2 p(x), \tag{1}$$

Mutual information measures the information shared between two random variables X and Y. Consider two random variables X and Y with joint probability distribution $p(x, y)$, and distribution of marginal probability $p(x)$ and $p(y)$. Thomas Cover [2] defined mutual information by the following equation 2:

$$I(X, Y) = \sum \sum p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \tag{2}$$

mutual information can be written in terms of entropy 3:

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \tag{3}$$

On the other hand, feature importance is a technique used to score the input features of a predictive model, indicating the relative importance of that feature concerning all in predicting the target variable. In this case, the decision tree was the technique used to determine feature importance, which offers a score based on the reduction of criteria using data division, criteria such as the Gini index or entropy, already mentioned here. The functions of the sklearn package were used for the selection of relevant attributes.

2.3 Modeling

After cleaning the data and exploratory analysis, the data is initially separated into training and testing. In the test part, 1000 samples of each category (pass and failure) are selected, totaling 2000 samples to simulate new data in the best-developed model. The rest of the data (not considering the data selected for testing) is used to develop the best model. It is noteworthy that the existence of duplicate data was evaluated to avoid leaks to the test portion.

Two techniques were applied in sequence for the balancing: oversampling and then undersampling. In the first technique, we sought to oversample the minority class to have 10 percent of the number of examples in the majority class, then used random subsampling to reduce the number of samples in the majority class to have 50 percent more than the minority class.

For the training validation, we evaluate the performance of the best hyperparameters. This part is necessary as it prevents overfitting to the final test portion. For the validation part, we choose 20% of the training base. The development of the best model followed the evaluation steps of 4 models: Decision Tree, Random Forest, Bagging with Decision Tree and Gradient Boosting.

3 Results

The exploratory analysis phase was essential for building new features with the help of the business area. This perception was useful and later validated with the use of mutual information and feature importance, which made it possible to considerably reduce the model's features without any loss of information.

Results on the validation portion of the models developed are presented in table 1 which shows the metrics of the models considered and, in bold, the best values obtained.

It can be observed in table 1 that the model obtained excellent results on the test portion, thus characterizing its high capacity to identify fraudulent transactions. It is important to highlight that the other models were also evaluated on the test portion, and, as expected, the results of the metrics did not stand out from the results of the best model (Bagging).

Table 1: Results of metrics on validation data. The best values are marked in bold.

Models	Accuracy	AUC	Precision	Recall	F1-Score
Decision Tree	0.9792	0.9785	0.9749	0.9785	0.9843
Radom Forest	0.9954	0.9955	0.9943	0.9955	0.9966
Bagging (DT)	**0.9962**	**0.9960**	**0.9954**	**0.9960**	**0.9972**
Gradient Boosting	0.9923	0.9923	0.9903	0.9924	0.9942

Another essential analysis to be done about the test portion is to evaluate the number of False Positives, which represent the number of transactions approved when they shouldn't be. Figure 1 presents the confusion matrix over the test data using the Bagging model.

The high capacity to identify fraudulent transactions is evident. Although the model does not reprove 34 samples, these represent a portion of only 3.4% of the total evaluated.

4 Conclusion

The presented model has excellent evaluation metrics for classifying expenses as fraudulent or not. Based on the classification project, it is possible to mitigate the number of occurred frauds in the expenses refund process by the employee or even reduce the errors of systemic approval flow expenses refund. Another important point in using the classification algorithm is that there is no gain with the refund audit process, which focuses on avoiding deviations and frauds, whether due to a lack of clear policies or bad faith. However, it has a direct contribution - the financial refund audit avoids losses that affect the company's management.

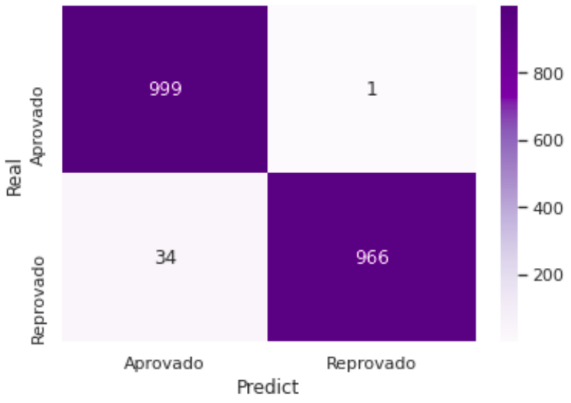

Figure 1: Confusion matrix with test data using the best model (Bagging)

Therefore, the gain obtained with the current model is related to the increase in revenues and an improvement in the operational refund process due to the loss avoidance, or even speed up errors treatment that may interfere in the expenses refund flow.

References

[1] SHANNON, C, E.A mathematical theory of communication. Bell System Technical Journal, v.27, n. 4 p.623-656, oct 1948. ISSN 00058580.

[2] COVER, Thomas M.; THOMAS, Joy A. Elements of Information Theory.Hoboken, NJ, USA: John Wiley & Sons, INcm, 2005. ISBN 9780471241959.