Data-driven Risk Assessment in Infrastructure Networks

António Pereira Barata[1,2], Gerrit-Jan de Bruin[1,2], Frank Takes[1], and Cor Veenman[1,3]

{a.p.pereira.barata,g.j.de.bruin,f.w.takes,c.j.veenman}@liacs.leidenuniv.nl
[1]Leiden Center of Data Science, Leiden University
[2]Human Environment and Transport Inspectorate, Ministry of Infrastructure and Water Management
[3]Data Science Department, TNO

1 Introduction

Governmental bodies commit to ensuring citizen safety, security, and trust. In order to make the highest impact possible, these agencies need to wisely choose how to allocate their limited assets. To accomplish this, the Human Environment and Transport Inspectorate (ILT) recognized the need for data science and network science as tools for data-driven approaches for their inspections. In this paper we describe a use case in which machine learning is able to identify possible fraudulent waste notices. The Waste Shipments Regulation (WSR) comprises the legislation that companies must follow in order to transport waste materials through a European Union (EU) member state. The legislation requires that a company wanting to transfer waste reports the type of waste, as the cost for each type is different. As a result, some companies might intentionally mislabel their waste. Hence, finding these mislabeled waste transports using machine learning techniques is the topic of this paper.

2 Data

ILT publishes all notices of waste transport on their website[1]. We obtained a dump of this data belonging to the years 2008 through 2016. The data contains 300 thousand entries of 19 variables. Because some of them were of low quality, only 9 nominal and 1 numerical variables were selected. Type of waste (target) is categorized in 20 classes defined in the European List of Waste (LoW)[2].

[1] https://english.ilent.nl/themes/international-shipment-of-waste

[2] http://ec.europa.eu/environment/waste/framework/list.htm

3 Problem statement

The goal of this study is to use anomaly detection to find transports that strongly deviate from other transports within the reported waste category, which could imply the fraudulent behavior described in Section 1.

4 Approach

We perform anomaly detection by means of supervised learning. A standard linear support vector machine classifier (LSVMC) was chosen to model our data [1]; a five-fold train-test split was performed for cross-validation. It is necessary for categorical variables to have their values represented numerically. Therefore, we performed variable binarization on all categorical variables. We model each category class (one-vs-rest) and determine, for every point labeled as that category, whether it is an outlier. An outlier is defined as a data point with a score below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$, where IQR (inter-quartile ratio) depicts the value difference between the first and third quartiles (Q_1 and Q_3, respectively), within all false negative (FN) scores associated to a distinct classifier. If a data point is determined as an outlier by a model with good performance, this could be indicative of non-compliance. As the size of the class categories varies significantly, evaluation metrics that take into account the number of true negatives (e.g., accuracy fall-out, ROC and AUROC) are not applicable. Instead, we will use average precision (AP): a high value indicates that the model is able to efficiently classify data points regarding a particular class. [2]. Given well-performing classifiers, the most extreme low scores are most probable outliers.

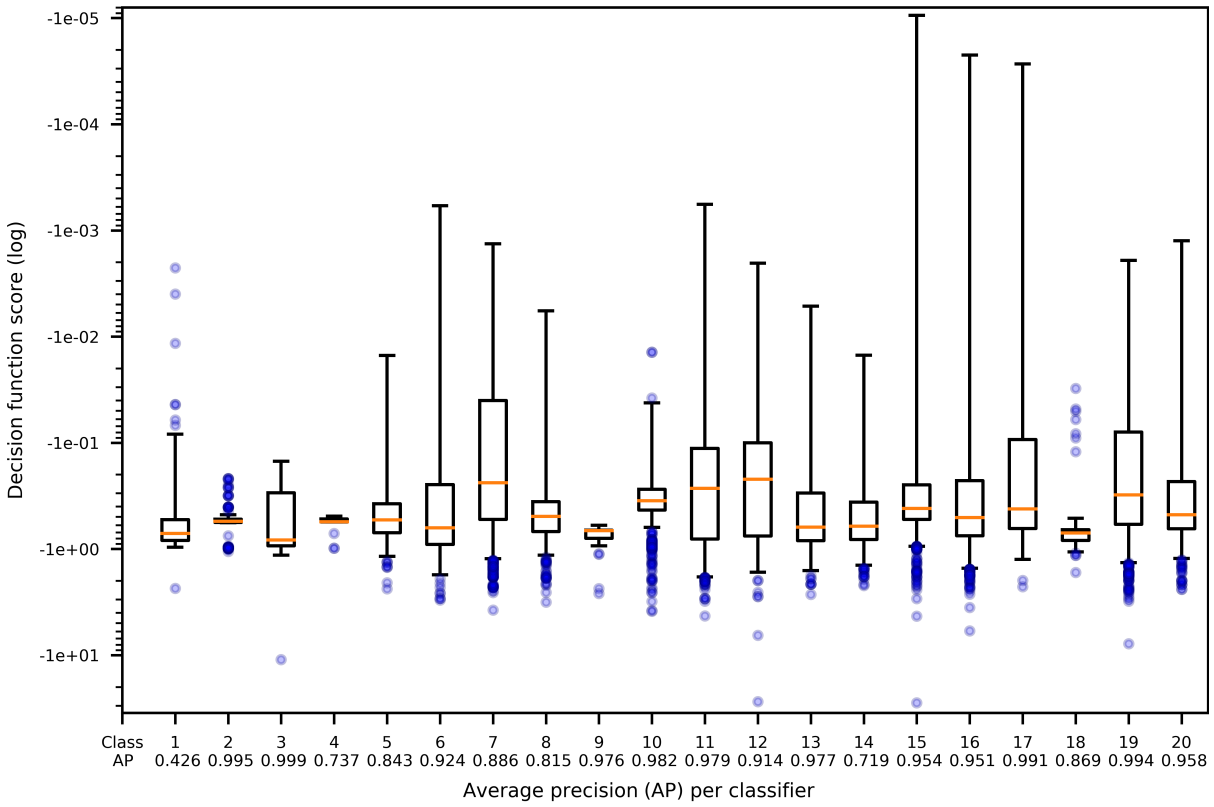

Figure 1: Visual representation (boxplot) of the distribution of decision function scores (log scale) belonging to FN predictions per unique classifier and associated average precision (AP) rounded to 3 decimal places.

5 Results

As shown in Figure 1, out of the total 20 classifiers, 13 exhibited an average precision score ≥ 0.9 while only 3 classifiers were evaluated at a score < 0.8. Considering classifiers with high AP value, each lowest-scoring FN prediction per category classifier is most likely an outlier: the data-point is furthest away from all other data points pertaining to the same class, within feature space. For example, the most reliable classifier (AP ≈ 1) is the support vector machine trained to determine whether a data point belongs to the third LoW category. One can expect, hence, that its predictions are reliable. By analysing the score distribution of classifier 3, it is simple to observe a data point with a score close to -10 (blue point immediately below the lower whisker of the corresponding boxplot). This means that the feature values associated to this data point are significantly different from all other points with the same label. Therefore, it is a data point that should be inspected further.

6 Conclusion

This work aims to determine what transportation entries are the most likely to be fraudulently documented, making use of data-driven approaches. In order to validate our results, we will cooperate with domain experts in future work to check whether the outliers predicted by our model are indeed associated to non-compliant transportation events and, consequently, company misbehaviour that may result in applicable sanctions.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2 ed., 2009.

[2] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS ONE*, vol. 10, no. 3, 2015.