Road Accident Predictions as a Classification Problem

Madhulika Agrawal
magrawal@uevora.pt
Teresa Gonçalves
tcg@uevora.pt
Paulo Quaresma
pq@uevora.pt

Departmento de Informática,
Universidade de Évora, Portugal

Abstract

This paper aims at evaluating the performance of various classification methods for road accident prediction. The data is collected under MOPREVIS [3] project which aims at improving road safety in Portugal. The data is highly imbalanced as there are fewer accident instances than the non-accident ones and due to this imbalance, it is observed that the traditional classification algorithms do not perform well. Using sampling techniques (undersampling and oversampling) improved the results but not significantly. Some methods resulted in increased recall but that decreased precision as the algorithm returned more false positives to make up for data imbalance.

1 Introduction

According to the annual report published by Autoridade Nacional Segurança Rodoviária (ANSR) [1], in 2020 there were 26,501 accidents with victims on the continental Portugal, which resulted in 390 fatalities, either at the scene of the accident or during transport to the hospital. There were 1,829 serious injuries and 30,706 minor injuries.

In the period from January 1st to March 18th 2020, before the first period of lockdown resulting from the first State of Emergency, there was a general reduction in accidents when compared to the same period in 2019: 424 fewer accidents (-6.2%), 22 fewer fatalities (-22.0%), 41 fewer serious injuries (-9.6%), and 536 fewer minor injuries (-6.5%). In global terms, compared to 2019, in 2020 there was an improvement in the main accident indicators: 9,203 fewer accidents (-25.8%), 84 fewer deaths (-17.7%), 472 fewer serious injuries (-20.5%) and 12,496 fewer minor injuries (-28.9%).

Although there were fewer accidents in 2020, they are still not zero. And it is important for administration of the country to make transportation as safe as possible. In our research, our aim is to identify the time and spot that are most susceptible to accidents. This will help improving the road safety.

We discuss the problem definition in the next section which is followed by related work. Description of the dataset and the experimental setup, along with result discussion can be found in Section 3 and Section 4 respectively. The paper is concluded with scope for future work.

2 Problem Definition

There are several factors that could cause a road accident. Over the years, researchers are investigating the impact of these numerous variables and their significance in causing road accidents.

Road accident prediction can be defined as a two-class classification problem: an accident on a road, at any given time constitutes the positive sample; all the other times (when there are no accidents) are negative samples. The final problem definition is to identify if there will be an accident at a specific place and time instant.

By framing accident prediction as a binary classification problem, a major restriction arises: very high class imbalance. The number of instances when accidents happen are much lower than the non-accidents. The dataset built, described in the next section, has only 0.022% positive samples. This creates bias for the negative class while training a machine learning algorithm. One of the possible ways to take care of class imbalance is to create a balance between positive and negative samples by undersampling or oversampling [4] before training the machine learning model.

3 Related Work

There is plenty of research on how to manage data imbalance. The survey paper by Ander Carreño *et al.* [5] summarizes various class imbalance studies. There are precisely three types of methods to handle data imbalance [7]; data-level methods, algorithm-level methods and hybrid methods. The use of Random Forest (RF) classifier along with sampling techniques for imbalanced data has been explored by Chao Chan *et al.* [6]. They tested their model on 6 different dataset with varying degree of data imbalance.

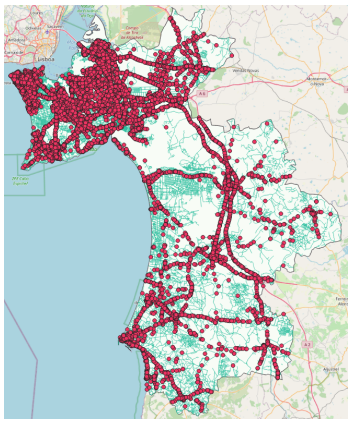

Figure 1: Distribution of Accidents in Setubal (PT)

Figure 2: Distribution of Accidents on N10 Road

4 Dataset

The accident database, provided by Guarda Nacional Republicana (GNR), consists of all the accidents that occurred in the Setubal region of Portugal between 1-January-2016 to 31-December-2019. The location coordinates of the accidents were mapped to the roads on Open Street Maps (OSM) [8]. For sake of simplicity, in our research we are focusing only on the accidents on National Road 10 (N10). Distribution of accidents in Setubal and N10 is presented in Figure 1 and Figure 2.

OSM divides the road into smaller, unequal segments. There are 186 road segments on N10 that had at least one accident in 4 years. The statistical analysis of number of accidents on various road segments and the length of those segments are presented in Table 1.

	#accidents	length (in meters)
mean	7.69	480.47
std	10.07	1027.19
min	1.00	21.21
25% percentil	2.00	82.29
50% percentil	4.00	186.87
75% percentil	9.00	464.29
max	71.00	8774.24

Table 1: Number of accidents per road segment and their length

For each road segment, the dataset built contains an entry for each hour of the day of the four years study. A sample is marked positive if an accident occurred in the one hour window for that day and is marked negative otherwise.

Along N10, Instituto Português do Mar e da Atmosfera (IPMA) have three weather stations. Depending on the proximity of the weather station to the road segment, the weather information for each instance was recorded from the nearest weather station. Other features, presented in Table 2, were also recorded and can be categorized as follows:

- **Time-Invariant Features:** Features of the road segment that do not change with time, like the presence of bridge or tunnel. There are 11 time invariant features;

- **Time-Variant Features:** Features that changes with time, like weather information. There are 12 time variant features associated with each segment of road.

	Feature	Description
Time Invariant	osm_id	unique ID of the road segment
	oneway	binary, if the road is a one-way
	bridge	binary, presence of bridge
	tunnel	binary, presence of tunnel
	codsubunidade	ID of the locality
	codpostal	postal code
	codconcelho	ID of concelho
	codfreguesia	ID of freguesia
	codsensepc	binary, if the road has central divider
	codtracado	binary, if the road is straight
	codtipoloc	binary, if the road is inside the city
Time Variant	ano	year
	ms	month
	di	day of the month
	hr	hour of the day
	day_of_week	sunday, monday, tuesday...
	daylight	dawn, morning, afternoon, dusk or night
	t_med	Average air temperature at 1.5m
	hr_med	average relative humidity
	dd_med	average wind direction
	ff_med	average wind intensity
	pr_dur_acc	accumulated precipitation duration
	pr_qtd_acc	accumulated precipitation quantity

Table 2: Features of the Dataset

There are features for which the values were missing for some hours; those data points we removed from the dataset. Table 3 details the size of the dataset, before and after removing data points with missing values.

	Unclean Data	Clean Data
Total Samples	6,521,904	6,402,059
Positive Samples	1427	1416
Negative Samples	6,520,477	6,400,643

Table 3: Description of the Dataset

5 Experimental Setup

Imbalance learn [2] is an open source library relying on Scikit Learn and provides tools for handling class imbalance in classification and includes implementation of various undersampling and oversampling methods. Before training a machine learning model, the samples are either under-sampled (from the majority class) or over-sampled (from the minority class) to create a balance between the two classes. We tested 10 undersampling, 5 oversampling and 2 joint under and over sampling methods on our data. The dataset is divided into train and test sets in the ratio of 70:30. All the methods are used with their default parameters. The performance of the algorithms are measured over AUC-ROC, Precision, Recall, and F1 Score.

	Algorithm	AUC	Prec	Rec	F1
RF	w/o Class Weights	.5190	**.1818**	.0381	.0631
	Balanced Class Weights	.6035	.0730	.2076	**.1080**
	Bootstrap Class Weights	.6035	.0726	.2076	.1076
Bal. RF	w/o Class Weights	.7364	.0005	.7780	.0011
	Balanced Class Weights	.7379	.0006	.7231	.0012
	Bootstrap Class Weights	**.7433**	.0005	**.7947**	.0011
	Easy Ensemble Classifier	.6939	.0004	.7589	.0008

Table 4: Results of Different Classifiers

The performance of different classification algorithms on the imbalanced dataset is presented in the Table 4. As can be seen, Balanced RF [6] outperformed all the other methods. This is why it is used as base classification algorithm for all the remainder of the experiments. The results of different sampling algorithms are presented in the table 5. Undersampling of the majority class gave better results than any other.

	Algorithm	AUC	Prec	Rec	F1
Oversampling	ADASYN	.5261	.0709	.0525	.0603
	BorderlineSMOTE	.5237	.0651	.0477	.0550
	RandomOverSampler	.6035	.0713	.2076	.1061
	SMOTE	.5249	.0677	.0501	.0576
	SVMSMOTE	.5702	**.0993**	.1408	**.1164**
Undersampling	AllKNN	**.7471**	.0005	.8019	.0011
	ClusterCentroids	.7289	.0005	.7780	.0010
	EditedNearestNeighbours	.7364	.0005	.7804	.0011
	InstanceHardnessThreshold	.7389	.0005	.7780	.0011
	NearMiss	.4937	.0002	**.9809**	.0004
	NeighbourhoodCleaningRule	.7438	.0005	.7923	.0011
	OneSidedSelection	.7361	.0005	.7708	.0011
	RandomUnderSampler	.7334	.0005	.7708	.0011
	RepeatedEditedNN	.7404	.0005	.7852	.0011
	TomekLinks	.7376	.0005	.7804	.0011
Joint	SMOTETomek	.5249	.0673	.0501	.0574
	SMOTEENN	.5321	.0868	.0644	.0739

Table 5: Results of Under and Over Sampling

Analysing the results, the following general observations can be made:

- **Low Precision:** Since there are more negative samples in the dataset hence there are higher chances of them being classified as false positives by the classifiers.
- **High Recall:** Fewer positive samples converts to fewer false negatives in classification and therefore higher recall.

6 Conclusion and Future Work

The higher recall by some methods is because the classifier is classifying more samples as positives which also results in increased false positives, causing reduced precision. From the results obtained, it is evident that portraying accident prediction as a classification task is not efficient. In future, it would be interesting to see what kind of results we could get by treating this as an anomaly detection problem, considering accidents as the abnormal behaviour in the regular road data.

7 Acknowledgment

This research is supported by the MOPREVIS - Modelação e Predição de Acidentes de Viação no Distrito de Setúbal, funded by Fundação para a Ciência e Tecnologia (FCT), reference FCT DSAIPA/DS/0090/2018 under the National Initiative on Digital Skills 2030, Portugal.

References

[1] Relatórios de sinistralidade. URL http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx.

[2] Imbalance learn documentation. URL https://imbalanced-learn.org/stable/.

[3] Moprevis. URL https://moprevis.uevora.pt/.

[4] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced distributions. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[5] Ander Carreño, Iñaki Inza, and Jose A Lozano. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, 53(5):3575–3594, 2020.

[6] Chao Chen, Andy Liaw, Leo Breiman, et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1-12): 24, 2004.

[7] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[8] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.