University of Milano-Bicocca

Machine Learning

Classification Analysis of Crimes in the Los Angeles' districts

Salma Chabib[1], *Lilia Grasso*[2], *Daniele Mingolla*[3], *Alice Schiavone*[4]

[1] *Bachelor's degree in Computer Science, University of Insubria*
[2] *Bachelor's degree in Physics, University of Milano-Bicocca*
[3] *Bachelor's degree in Computer Science, University of Perugia*
[4] *Bachelor's degree in Computer Science, University of Insubria*

Abstract: Living in a big city as Los Angeles comes with its pros and cons, like being exposed to a higher number of crime episodes. To help its citizens or tourists preparing themselves for the worst case scenario, this paper investigates the possible classification techniques that could be used to inform people of the most probable type of crime incident. Starting from a huge database, it's possible to extract anonymous information about past incidents' victims and use this data to predict which crime one could suffer from based on its personal characteristics or area of transit. After pre-processing the data, models can be trained on a subset of records of the original data set and evaluated on the complementary subset. Analyzing the different performance measures obtained from different methodologies, it is possible to say that the best approach to this particular classification problem -i.e. the model that presents a higher value of recall, is the *J48 model*. We aimed to obtain an high value of *Recall*, because, considering a victim's point of view, she/he would prefer to receive a false alarm instead of underestimating the situation.

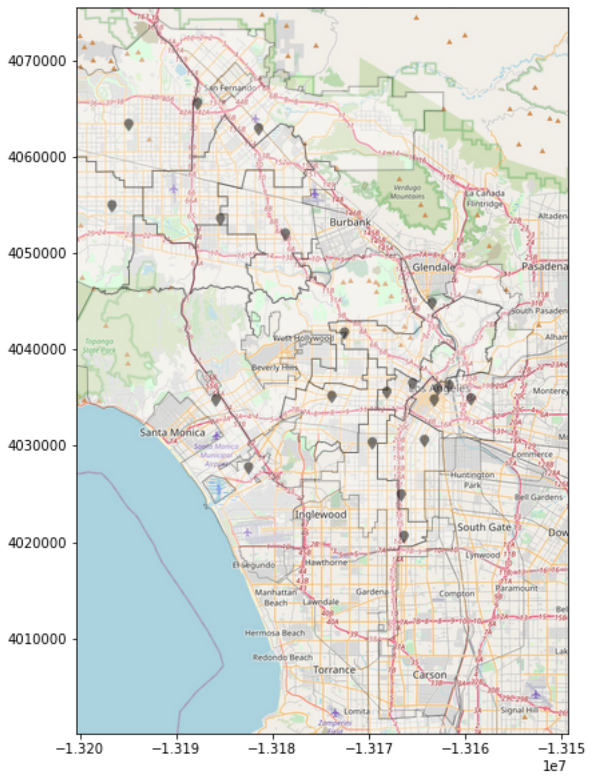

Fig. 1: Map of Police Departments in Los Angeles' districts (LAPDs)

1 Introduction

The City of Los Angeles is among the most populous cities in the United States, second only to New York, and its 4 million residents live in an area that covers more than 1200 square kilometers [3]. We are interested in training a model that correctly predicts what type of crime incidents one could encounter visiting or living in the city, based on the characteristics of such a person. Doing so, people could be more careful to some type of felony instead of another. Given that caution is always advised against any crime, this model could still be an interesting instrument. To provide an overview of the Los Angeles' districts it has been plotted, in figure 7, a map of the districts and the police departments [2] per each district (the black reverse drop in the map), showing the vastness of the area considered. The crime reported by the local authorities have a distinctive code, based on the type of crime committed. The original data set contains a large number of identification codes, that cover all the possible shades of an incident. The map in figure 2 represents the four main categories of crimes occurred in October of 2019 in the districts of Los Angeles.

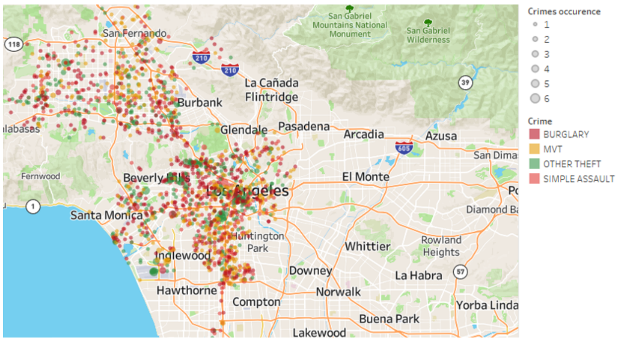

Fig. 2: Occurrences of crimes at October 2019 in Los Angeles' District.

It's worthwhile to notice that at a first glance the districts that have suffered from most crimes in October are in Central, Southeast and Southwest Los Angeles. We aim to identify which crimes it is possible to suffer from in the Los Angeles' streets and this problem can be clearly identified as a multi-class classification problem, because the value of the attributes to be predicted are mutually exclusive and they are more than two. This approach raises the following questions: how can a such large number of attributes be handled correctly by a classification model? How to act efficiently when dealing with such a high number of observations (more than 2 millions)? And most importantly, in this particular case, which models perform better than others?

1.1 ***Data Set Description***

To monitor crime incidents, the Los Angeles Police Department (LAPD) transcribes data directly from crime reports to the online open data set catalogue [1] of the city. The 'Crime Data from 2010 to 2019' gives details about each felony: each row represents (anonymously) a crime incident, from its location to the type of incident. The data set is obviously imbalanced, as the number of reported crimes in each category is naturally very different: the number of violent crimes are a very small minority compared to the number of mild felonies.

1.1.1 ***Attributes Description****:*

• **AREA**: numerical attribute. The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1 to 21.
• **Rpt Dist No**: numerical attribute. It's a four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons.
• **Crm Cd**: numerical attribute. Indicates the crime committed, it's equal to Crm Cd 1.
• **Vict Age**: numerical attribute indicating the age of the victim. When the age was unknown, the value 0 has previously been assigned by who filled in the dataset.
• **Vict Sex**: categorical nominal attribute. The tokens are F - Female, M - Male, X - Unknown.
• **Vict Descent**: categorical nominal attribute. It describes the ethnicity of the victim. The tokens are: A - Other Asian, B - Black, C - Chinese, D - Cambodian, F - Filipino, G - Guamanian, H - Hispanic/Latin/Mexican, I - American Indian/Alaskan Native, J - Japanese, K - Korean, L - Laotian, O - Other, P - Pacific Islander, S - Samoan, U - Hawaiian, V - Vietnamese, W - White, X - Unknown, Z - Asian Indian.
• **Premis Cd**: numerical attribute. It refers to the type of structure, vehicle, or location where the crime took place.
• **Weapon Used Cd**: numerical attribute. It refers to the type of weapon used in the crime.
• **Weapon Used Desc**: categorical nominal attribute. It defines and describes the Weapon Used Code provided.
• **DR_NO**: division of Records Number; is an official file number made up of a 2 digit year, area ID, and 5 digits.
• **DATE_RPTD**: the date when the crime has been reported to the police station.
• **DATE OCC**: date when the crime has occurred.
• **TIME OCC**: time when the crime has occurred.
• **PART 1-2**: numerical attribute. The crimes has been split into two main categories the brutal ones and the non brutal ones.
• **AREA NAME**: categorical nominal attribute. The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.
• **CRM CD DESC**: categorical nominal attribute. It defines and describes the Crime Code provided.
• **MOCODES**: numerical attribute. It's the Modus Operandi: activities associated with the suspect in commission of the crime.
• **PREMIS DESC**: categorical attribute. It defines and describes the Premise Code provided.
• **STATUS DESC**: categorical nominal attribute. It defines the Status Code provided.
• **CRM CD 1**: numerical attribute. It indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.
• **CRM CD 2**: numerical attribute. May contain a code for an additional crime, less serious than Crime Code 1.
• **CRM CD 3**: numerical attribute. May contain a code for an additional crime, less serious than Crime Code 1.
• **CRM CD 4**: numerical attribute. May contain a code for an additional crime, less serious than Crime Code 1.
• **LON**: numerical attribute. Longitude.
• **LAT**: numerical attribute. Latitude.
• **CROSS STREET**: categorical nominal attribute. Cross Street of rounded Address.
• **LOCATION**: categorical nominal attribute. It's the street address of crime incident rounded to the nearest hundred block to maintain anonymity.
• **Status**: categorical nominal attribute. It describes the status of the case. (IC is the default)

2 Pre-Processing

Data pre-processing is an important step in Machine Learning because the quality of data and the information that can be derived from it, directly affects the ability of a model to learn. We considered only the most recent crimes from 2016 to 2019 because:

1. being recent, they could more representative of future event;
2. the high dimension of the data set incredibly slowed down the process of data mining;

Furthermore, the attributes deemed not useful for our purposes, the irrelevant ones, have been removed. In the data set there are categorical attributes that we have decided not to consider, because each of them corresponds to a unique number stored in another attribute though they can be labelled as redundant attributes. In such case, to facilitate the training of the models, we considered only numerical attributes, specifically:

• Crm Cd instead of Crm Cd Desc and Crm Cd 1.
• Premis Cd instead of Premis Desc.
• Weapon Used Cd instead of Weapon Desc.

Finally, we have decided to use the following attributes for our analysis:

• AREA;
• Rpt Dist No;
• Crm Cd;
• Vict Age;
• Vict Sex;
• Vict Descent;
• Premis Cd;
• Weapon Used Cd;
• Status.

2.1 ***Dimensionality Reduction & Attributes Selection***

Because of the large number of crime types, we decided to aggregate some of them to obtain four macro-categories, excluding less frequent and/or violent crimes:

• **Simple assault** that identifies lynching or lynching attempted, resisting arrest, battery on Firefighter, battery on Police Officer, spousal abuse, child abuse, throwing substance at vehicle, stalking, threatening phone calls or letters and criminal threats;
• **MVT** that identifies stolen vehicles and stolen vehicles attempted;
• **Burglary** that identifies burglary or burglary attempted;

- **Other theft** that identifies theft, shoplifting, dishonest employee, till tap, theft from coin, theft of a bicycle or a boat (attempted or stolen).

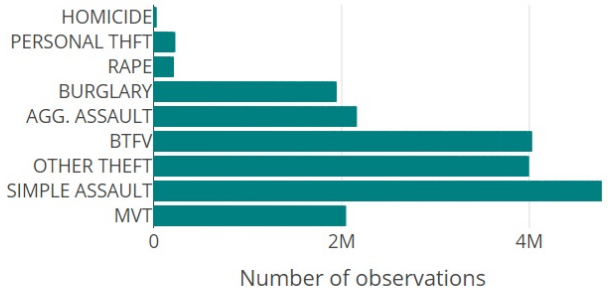

Fig. 3: Composition of a sample of the original data set, by crime macro-categories

Using this aggregation we reduced the number of classes (initially more than 10) to be predicted, furthermore, we removed from the data set those crimes whose category did not fit into any of those previously listed. In order to increase interpretability, to facilitate graphical representation, to reduce the amount of time and memory and more important to allow data mining algorithms to work better, we applied dimensionality reduction to lower the number of attributes deleting ones that we considered irrelevant:

- DR_NO
- DATE OCC
- DATE_RPTD
- TIME OCC
- PART 1-2
- AREA NAME
- CRM CD DESC
- MOCODES
- PREMIS DESC
- STATUS DESC
- CRM CD 1
- CRM CD 2
- CRM CD 3
- CRM CD 4
- LON
- LAT
- CROSS STREET
- LOCATION

This step is crucial in order to avoid the curse of dimensionality referring to a set of problems that arise when working with high-dimensional data. The dimension of a data set corresponds to the number of attributes/features that exist in a data set. Some of the difficulties that come with high dimensional data manifest themselves either during the analysis or the visualization of the data to identify patterns or training machine learning models.

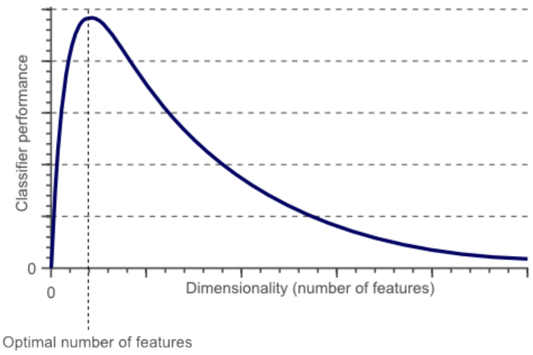

Fig. 4: Curse of Dimensionality

2.2 *Equal Size Sampling*

After consulting the literature [5] we performed stratified undersampling that does not keep the proportions of the original data set, in fact it includes 16168 observations for each crime type. This particular number is due to the node performing equal size sampling, because it adapts to the smallest number of occurrences of the target variable, therefore we decided to exclude the less frequent macro-categories in order to not have a limited data set. The equal sampling operation is useful to avoid classification problems due to class imbalance. The final sample obtained consists of 248356 observations.

2.3 *Handling with Missing Values*

The majority of attributes that we have decided to keep in our analysis contain missing values and in order to avoid problems in the development of predictive models, we handled them in the following way: we applied the technique of record removal to "Crm Cd", "Status" and "Premis Cd" variables since the number of rows in which these attributes are NAs was very small; while the remaining missing values were managed as follows:

- Vict Age: replacement with the most probable value;
- Vict Sex: replacement with a global constant (X) and record removal for rows whose value for the variable is equal to N or H;
- Vict Descendant: replacement with a global constant (X);
- Weapon Used Cd: replacement with a global constant (0);

To handle missing values that are in the attribute "Vict Age", we used the linear regression model. First, the data set has been split into two data sets: one whose rows that had value 0 have been excluded and the other where the 0-values have been included. On the data set where the 0-values were excluded we have trained the regression model and after the training we used the model to predict the value of the observations, whose value was 0, in the second data set. Finally, we have concatenated the two subsets, precisely the one that excludes 0 and the one with the predictions got from the model, obtaining a data sets with no missing values.

3 Models and Performance Measures

The following 5 models for predicting the value of *Crm Cd* have been evaluated:

- **Decision Tree Learner** (heuristic): it classifies the records by sorting them down the tree from the root to a leaf/terminal node that provides the classification of the records. Each node in the tree acts as a test case for the attribute Crm Cd, and each edge descending from the node corresponds to the possible answers (possibles crimes) to the test case.
- **J48** (heuristic): It is another implementation of a decision tree developed by the WEKA project team. The additional features of J48 are: accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc.
- **Random Forest** (heuristic): It is another implementation of a decision tree but while growing the trees it adds additional randomness to the model.
- **Naive Bayes** (probabilistic): It exploits Bayes formula and compute posterior probability of the class attribute, then it labels the records with the class value that maximizes the posterior probability.
- **Logistic** (regression based): It computes the posterior probability of the class attribute given the value of the explanatory attributes.

Each model was trained using three different approaches such as Holdout, Iterated Holdout and Cross Validation.

3.1 *Performance Measures*

To evaluate the performance of each classification model, it is needed to select a specific instances of a classification model capable to ensure the maximum possible predictive accuracy [4]. To compare the models we used:

3.1.1 Accuracy:

it measures the capability of the model to give accurate prediction on records that were not available when the model was developed. Considering D_T the training set of t records and D_{TS} the test set of v records, such that

$$D = D_T \cup D_{TS}, D_T \cap D_{TS} = \emptyset, m = t + v,$$

and being y_i the class value associated with the instance $x_i \in D_{TS}$ and $f(x_i)$ the class value for the same instances but predicted by the model, it can be computed the *accuracy* as:

$$acc(D_{TS}) = 1 - \frac{1}{v}\sum_{i=1}^{v} L(y_i, f(x_i))$$

where the function $L(y_i, f(x_i))$ is the loss function that it is equal to 1 if $y_i = f(x_i)$, while it is equal to 0 if $y_i \neq f(x_i)$. The accuracy treats each class as equally important so it's not optimal for class imbalance problem. Considering the confusion matrix that gives an overview of the records correctly and incorrectly classified, it can also be computed as:

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$

where TN is the number of negative records correctly classified, TP is the number of positive records correctly classified, FP is the number of negative records erroneously classified as positive and FN is the number of positive records erroneously classified as negative.

3.1.2 Recall:

it measures the fraction of positive records correctly predicted by the classification model. If the *Recall* is large then very few positive records have been misclassified as negative class (the number of false negative is low).

$$Recall = r = \frac{TP}{TP + FN}$$

3.1.3 Precision:

it measures the fraction of records that actually turns out to be positive among all the records that the classification models has classified as positive class. If the *Precision* is high, then the number of false positive is low.

$$Precision = p = \frac{TP}{TP + FP}$$

3.1.4 F-measure:

it summarizes *Recall* and *Precision* doing the harmonic mean between them. If the *F-measure* is high, then both of the metrics (*Recall* and *Precision*) are reasonably high.

$$F - measure = \frac{2rp}{r + p}.$$

We aim to obtain an high value of *Recall* in order to achieve a low number of false negative records of the attribute *Crm Cd*. This decision has been done primarily considering a victim's point of view making her/him easier to choose where to go for a walk or where to buy a house, mainly because she/he would prefer to receive a false alarm, in order to do not underestimate the situation.

4 Analysis and Results

Different methodologies have been used to obtain different data set partitions, in order to compare, once the models have been tested, the diverse results. Each partitioning operation of the initial data set was performed using the same random seed in order to obtain reproducible results. In the following tables the most significant results are highlighted with the teal color.

4.1 Holdout

The first approach, called Holdout, was implemented as follows: the entire data set was split into 2 partitions (Partition A and Partition B): the first one consists of 67% of the entire data set -i.e. equivalent to two third of data- and it represents the training set; the second one that corresponds to one third of the data set, it represents the test set. In this way the models were trained using the training set and successively validated using the test set. Secondary operations were made in order to perfect the metrics mentioned above, in fact, since we have obtained the performance measures for every model, it has been necessary to compute the mean for each metric. Diving deeper, Holdout consists of limiting the data which are used to learn the classifier, -i.e. some records are saved to estimate the reliability level of the classifier. Since the accuracy computed using Holdout procedure depends on only one third of the entire data set and it could be affected by bias, more robust estimates are required and they could be obtained using an Iterated Holdout or a Cross Validation approach.

Table 1 Results obtained by Holdout

Model	Accuracy	Recall	Precision	F-Measure
Decision Tree	0.834	0.834	0.833	0.833
J48	0.862	0.862	0.864	0.862
Random Forest	0.857	0.858	0.860	0.857
Naive Bayes	0.736	0.736	0.744	0.715
Logistic	0.736	0.752	0.743	0.746

4.2 Iterated Holdout

This step is similar to the previous one, but instead of splitting the data set into train and test set once, we have performed this procedure twice and for each iteration we saved the results of the respective models. The final metrics obtained for each model correspond to the average of the metrics obtained in each iteration. Assuming that we've assigned an index to each of the 5 models, in the following formulas we will use the symbol K to denote the $k - th$ model.

$$Accuracy_k = \frac{1}{2}\sum_{i=1}^{2} Accuracy_{i,k} \quad k \in [1, 2, 3, 4, 5]$$

$$Recall_k = \frac{1}{2}\sum_{i=1}^{2} Recall_{i,k} \quad k \in [1, 2, 3, 4, 5]$$

$$Precision_k = \frac{1}{2}\sum_{i=1}^{2} Precision_{i,k} \quad k \in [1, 2, 3, 4, 5]$$

$$F\text{-}measure_k = \frac{1}{2}\sum_{i=1}^{2} F\text{-}measure_{i,k} \quad k \in [1, 2, 3, 4, 5]$$

Compared to the classical Holdout approach, the Iterated Holdout allows to reduce the bias but requires more computational resources.

Table 2 Results obtained by Iterated Holdout

Model	Accuracy	Recall	Precision	F-Measure
Decision Tree	0.833	0.832	0.832	0.832
J48	0.860	0.860	0.863	0.860
Random Forest	0.855	0.855	0.857	0.854
Naive Bayes	0.737	0.736	0.745	0.715
Logistic	0.752	0.752	0.743	0.746

4.3 *Cross Validation*

The third approach utilized is called Cross Validation, more precisely 5-folds Cross Validation. The original data set, which was the output of the pre-processing phase, was partitioned in 5 exhaustive disjoint subsets, the same number of learning testing iterations done. During each iteration four of the subsets are used in order to train the model while the fifth one was hold out in order to test the model. In general, starting from K-folds, to each subset is given the opportunity to be used in the hold out set 1 time and used to train the model $K-1$ times; in this particular case: each subset was used four times as training set and one time as test set. Since we've obtained for each iteration the metrics' values, secondary operations were made in order to compute the metrics' mean both for each fold and model, so that we could have had clearer understanding of the model's performances. It is also good practice to include confidence intervals. It's worthwhile to notice that Cross Validation procedure ensures that each record of the data set is included into the training sets the same number of times and exactly one time in the test set. This methodology has led to less biased and less optimistic estimate of the model skills than other models but it's computationally expensive.

Table 3 Results obtained by Cross Validation

Model	Accuracy	Recall	Precision	F-Measure
Decision Tree	0.832	0.833	0.833	0.833
J48	0.861	0.861	0.863	0.860
Random Forest	0.856	0.856	0.859	0.856
Naive Bayes	0.736	0.736	0.745	0.715
Logistic	0.750	0.750	0.741	0.745

4.4 *Comparing Classifiers*

4.4.1 *Validation and Confidence Interval:*

We used the K-folds cross validation procedure to test whether or not there was a significant difference between the Accuracy of each model calculated on the same test set. The data set is partitioned into K disjoint subsets, exhaustive and with almost a constant number of records. The difference between their error rates during the $k-th$ iteration (fold) can be normally distributed and written as:

$$d_k = e_{1k} - e_{2k}$$

The overall variance in the observed differences is estimated using the formula on the left, while the mean of the difference between error rates during the $k-th$ iteration is estimated using the formula on the right:

$$\hat{\sigma}^2_{d^{cv}} = \frac{\sum_{k=1}^{K}(d_k - \bar{d})^2}{K \cdot (K-1)} \qquad \bar{d} = \frac{1}{K}\sum_{k=1}^{K} d_k.$$

We used the t-Student distribution to compute the confidence interval for the value of the true mean d_t^{cv}:

$$(\bar{d} - t^{K-1}_{1-\alpha/2} \cdot \hat{\sigma}_{d^{cv}}, \bar{d} + t^{K-1}_{1-\alpha/2} \cdot \hat{\sigma}_{d^{cv}})$$

We decided to calculate confidence intervals only for pairs of models that performed significantly better than the others. Thus, we have 3 models (Decision Tree, Random Forest, J48) for a total of 3 pairs. Table 4 shows that 0 is included in all of the calculated intervals with a 90% confidence level, so the differences among models are not statistically significant.

Table 4 Confidence Interval obtained for each model

Models Comparison	Lower Limit	Mean	Upper Limit
J48 & Decision Tree	-3.013	-0.028	2.957
J48 & Random Forest	-1.22	-0.004	1.211
Decision Tree & Random Forest	-2.543	-0.024	2.495

4.4.2 *ROC curves:*

The Receiver Operating Characteristic Curve is a graphical plot that illustrates the performance of the classifier without regarding to class distribution or error cost. Thus it provides a tool that is able to select optimal models and to discard non-optimal ones. The curve is formed by plotting on the x-axis the false positive rate (FPR) and on the y-axis the true positive rate (TPR).

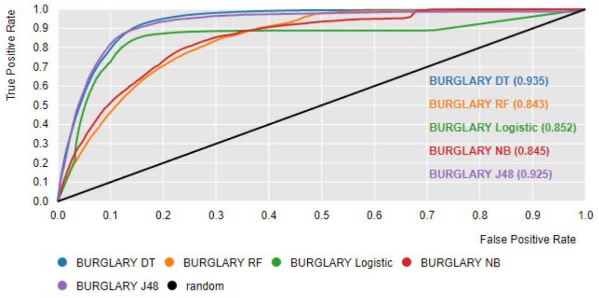

Fig. 5: ROC Curve of burglary crimes

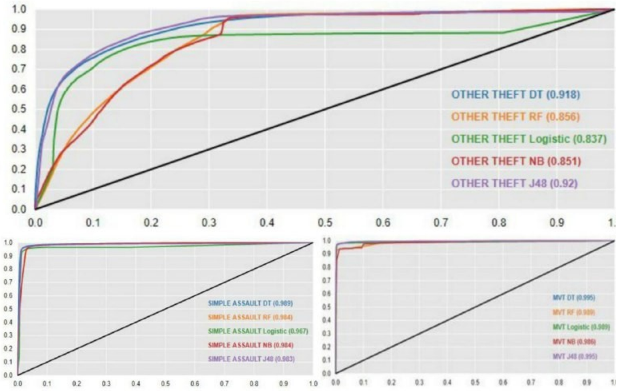

Fig. 6: Highlight of ROC Curves of the other crime categories. The first ROC is for "Other theft" crimes, the second ROC in bottom left is for "Simple Assault" crimes and the last one in the bottom right part is for "MVT" crimes.

All else being equal, we prefer a model who is capable to reach high value of True Positive Rate. In plot 5 the ROC Curve, regarding Burglary, illustrates that the J48 excels if the range of FP vary from 0% to 11% -i.e. on the left-hand side of the graph, covering about 87% of the true positive records. Clearly, if we aim to reach higher rate of TP than the other models presented for a range of %FP larger than 11%, the Decision Tree model should be chosen. To compare different classifiers, it can be useful to summarize the performance of each classifier into a single measure. One common approach is to calculate the area under the ROC curve, abbreviated to AUC (Area under the ROC Curve), that are presented in the plot in the brackets near the name of the model. It is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. It's worthwhile to notice that the black line represents the ZeroR model, based on random sampling. In figure 6 we have highlighted the ROC Curve of the other three macro-categories: Other Theft, MVT and Simple Assault.

Table 5 AUC per crime category

Model	Burglary	Other Theft	MVT	Simple Assault
Decision Tree	0.924	0.922	0.995	0.984
J48	0.935	0.917	0.995	0.989
Random Forest	0.853	0.838	0.988	0.967
Naive Bayes	0.846	0.851	0.987	0.984
Logistic	0.843	0.855	0.990	0.984

4.5 Feature Selection

The goal of Feature Selection is to find the best set of features that allows to build optimal models of the crimes' prediction. We designed and applied a Feature Selection procedure to discover which of the previous attributes are:

- Redundant: contains information already available;
- Irrelevant: contains information not useful to solve the considered data mining task;

Thus they can be removed without incurring much loss of information. We have opted for a Filter approach according to which attributes are selected before learning the classifier. In particular, we have implemented a multi-variate filter using the method CfsSubsetEval (that investigates on the correlations among attributes) finding the best subset whose explanatory attributes are not correlated among them but are correlated to the class attribute. This method has been applied on the models that didn't perform well as the tree-based classifiers, which naturally implement a Feature Selection approach when learning from the data (the feature selection is embedded). To investigate the performance of the Logistic model and the Naive-Bayes model, a filtering on their columns has been applied. This has led us to state that the following attributes are the more relevant ones to the classification problem for the models mentioned above:

- Vict Sex;
- Premis Cd;
- Weapon Used Cd.

A good subset of variables is constituted by those that are strongly linked to the class attribute but uncorrelated among them. The advantages achieved by Feature Selection are several, such as:

- the reduction of the cost of data collection;
- the reduction of inference time;
- the increase of interpretability;
- the increase of accuracy.

Table 6 Final results (Recall)

Model	Holdout	Iter. Holdout	Cr. Validation	Feat. Selection
Decision Tree	0.834	0.832	0.833	-
J48	0.862	0.860	0.861	-
Random Forest	0.858	0.855	0.856	-
Naive Bayes	0.736	0.736	0.736	0.762
Logistic	0.765	0.752	0.750	0.694

In this graph are plotted, per each classifier and after feature selection, the values of the recall: Comparing the results with the cross validation approach, the Logistic model performed better after filtering the data set attributes, but it is not the case for the Naive Bayes model, that significantly lost points in recall after the feature selection process. In conclusion, even after processing the data to facilitate the weakest models, the tree-based models are still the best performing models.

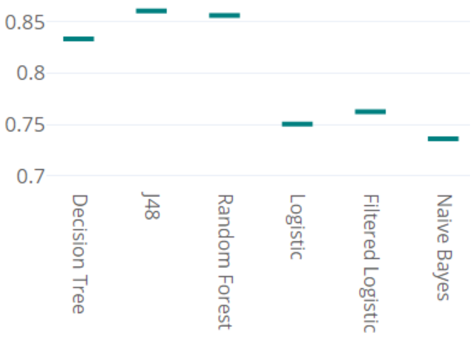

Fig. 7: Comparing classifiers by recall (after feature selection)

5 Conclusion

Following the various stages of evaluation, it is possible to say that to correctly classify a multiple number of classes of crimes, from a data set with a huge number of observations, it is best to chose tree-based classifiers (from the set of tested models). The features that influence the classification output are the victim sex, the location in which the crime took place and the (eventual) weapon used. There is not a significantly difference from one approach to another but the model that performs better is the J48 mainly for two reasons:

- J48 have the best results obtained by all the approaches (Holdout, Iterated Holdout, Cross Validation) in all types of performance measures;
- J48 have the largest Area Under the ROC Curve (AUC) in all crime categories, except for 'Other Theft', where it is still the second largest area.

Even though the results are encouraging, they can be improved by using more complex models such as Artificial Neural Networks or other different methodologies for example in the treatment of missing values. The implemented models do not take into consideration the eventual cost of classification, that could decrease the specific types of classification. This information was not available, but a special commission from LAPD could help data scientist train the model by estimating the cost of classification for each predicted record. Doing so, it would be possible to efficiently concentrate the training on a specific aspect of the process. Taking into account the similarity between the results obtained by the various models, it is useful to think about a study related to training-s speed of the models as the available data increases. In this way we can focus on the model that is able to update itself faster, allowing a saving in terms of time to the police and greater confidence in the predictions. These tools could be useful in a city like Los Angeles and also in any other big or small residential center, in fact they could help to raise awareness of the high quantity of crime incidents and they may also help police forces in detecting such crimes.

6 References

[1] 'Los Angeles Open Data', https://data.lacity.org/

[2] 'LA County: LA County-Sheriff and Police Department', https://public.gis.lacounty.gov/public

[3] Los Angeles: Geography and Climate', City-data.com (byAdvameg, Inc.), https://www.city-data.com/us-cities/The-West/Los-Angeles-Geography-and-Climate.html

[4] Data Mining-Evaluation of Classifiers, Institute of Computing Sciences, Poznan University of Technology, http://www.cs.put.poznan .pl/jstefanowski/sed/DM-4-evaluatingclassifiersnew.pdf

[5] The Proposal of Undersampling Method for Learning from Imbalanced Datasets, https://www.sciencedirect.com/science/article/pii/ S1877050919313456