MLDM Mini Projekt

Kim Lan Vu, Asha Schwegler, Paul Müller

18. Dezember 2022

Inhaltsverzeichnis

1 Introduction

1.1 Goal of the assignment

The goal of this assignment is to predict the type of crime from specific information like time and place. The authors of this paper analysed the data from crime reports from the city of San Francisco provided by the site "Kaggle".

2 Loading the Data

2.1 Getting the data

The San Francisco crime report was loaded into the notebook. At first glance the authors noticed that there are nine columns and 878050 data samples.

3 Processing and cleaning the Data

3.1 Processing and Cleaning the Data

In order to make predictions the authors decided that the PdDistrict would suffice as location information, so the address could be dropped. The description and resolution of the crime and arrest were also unnecessary to make a description of the type of crime. The authors changed the date into numbers and splitted into separate columns consisting of Year, month, day, hour and DayOfWeek.

While plotting the cooridinates, some outliers were noticeable.

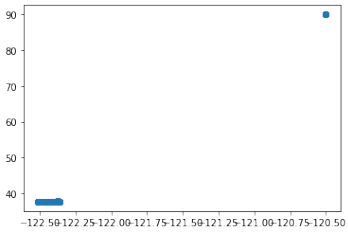

Abbildung 1: Plot of Cooridinates.

Since San Francisco has the latitude of circa 37, all coordinates with latitude under 40 were kept and the rest deleted, there were only 76 coordinates to be deleted.

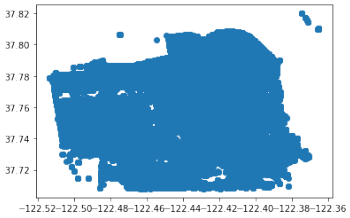

Abbildung 2: Plot of Cooridinates cleaned up.

3.1.1 Binning

To make the patterns more noticeable the authors decided to use Data binning. The Hours were binned in an equal-width size 6 bin, the divisions represent Early morning, morning, noon, afternoon, evening and night. Additionaly the longitude and latitude were binned separately in size 80 bins.

4 Plotting and analysing Data

4.1 Plotting and analysing the data

To get a bit of an overview of the situation the authors decided to create some plots.

4.1.1 Plot by Day of the week

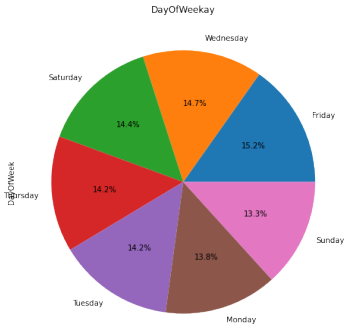

Abbildung 3: DayDistribution.

4.1.2 Plot by District

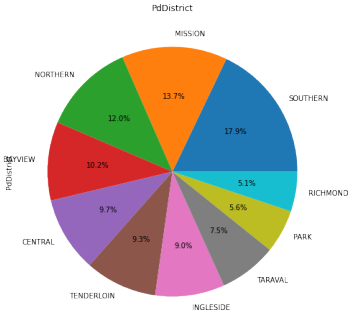

Abbildung 4: DistrictDistribution.

4.1.3 Plot by the hour

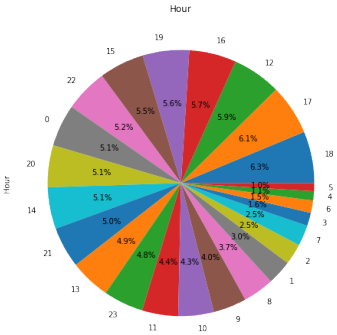

Abbildung 5: HourDistribution.

4.1.4 Map of districts

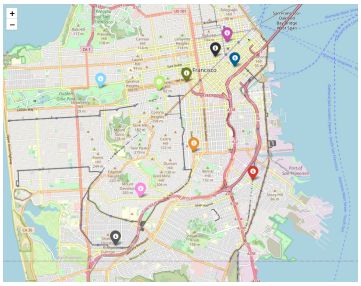

Abbildung 6: MapOfDistricts.

4.1.5 Plot by Year

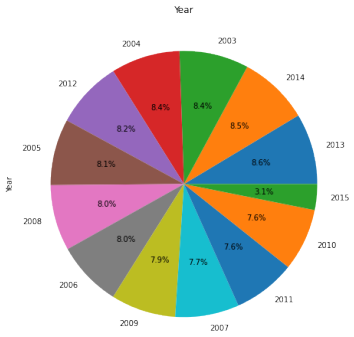

Abbildung 7: YearDistribution.

4.1.6 Plot by category

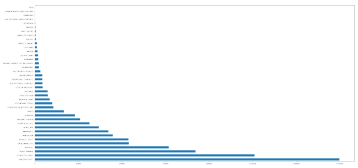

Abbildung 8: CategoryDistribution.

4.1.7 Conclusion

As one can see the crimes are relatively evenly distributed throughout the districts and dates. The only difference one can see is it depends on the hour of the day what the frequency of crimes are. Also the category samples are unbalanced.

5 Building the Models

The authors have tried the following models:

- SVM
- RandomForest
- NN

From all the models the Random Forest seemed to get the best score. With the following hypertunings due to hardware restrictions(e.g crashing due to overuse of RAM capacity) and for better log loss scores :

- n_estimators:525
- max_depth:15
- max_features:sqrt
- min_samples_leaf:15
- min_samples_split:30
- random_state:0xdeadbeef
- verbose:1
- n_jobs:4

The final score this recieved was:

Abbildung 9: FinalScore.