World Data League

Insights Report 2021

Authors: Leonid Kholkine and Miguel José Monteiro
Editor: Fabiana Oliveira
Designer: Celso Santana

The following have contributed directly or indirectly to the content of this report:

WDL Participants: Ahmed Fares, Alberto Mesquita Santos, Alessandro Padella, Alessio Arcudi, Alex Ablanya, Alexander Obenauff, Alvaro Ponte Blanco, Ana Almeida, Ana Carvalho, Ana Gonçalves, Ana Silva, Angelo Figueiral, Ankit Chaturvedi, Begoña Echavarren Sánchez, Beltran Vazquez, Bharathi A, Bruno Silva, Carolina Alves, Carolina Bellani, David Martins, Débora Martins, Diego García Lozano, Diogo Nasser, Diogo Seca, Fábio Tomaz, Francesco Vinci, Francisco Cristóvão, Giulio Mazzanti, Gustavo Fonseca, Håkon Sandaker, Hardique Manilal, Imre Boda, Isabel Chaves, Ishan Mistry, Jay Mundhara, Jedrzej Alchimowicz, Jelilat Anofiu, Joana Godinho, Joana Morgado, João Gamboias, João Louro, Johanna Ott, Jorge Gómez Berenguer, Jorge Gomez-Lechón Verdia, José Mota, Kaushik Valluri, Kiran Karthika Divakar, Krishna Verma, Kristof Menyhert, Kylix Alexander da Silva Alves Afonso, Laszlo Szilagyi, Leonardo Santangelo, Lilla Tagai, Lorenzo Barcella, Lúcia Moreira, Manuel Pertejo Lope, María Javierre, Mariana Monteiro, Martin Hadid, Matteo Pozzan, Michele Nasini, Moritz Geiger, Nicholas Sistovaris, Nidhi Pusadkar, Nirbhaya Shaji, Nissimol Aji, Nuno Lavado, Paulo Maia, Pedro Araújo, Pedro Faria, Pedro Miguel Rolim, Prasath M, Raquel García Meneses, Ricardo Araújo, Ricardo Azevedo, Rui Granja, Sachiyo Daley, Safia E K, Sai Pravallika Myneni, Sandra Martínez Sanchis, Saurabh Joshi, Sebastian P Joseph, Semanur Kapusizoğlu, Sergio Calderón Pérez-Lozao, Shubham Gandhi, Shubham Jangir, Sijuade Oguntayo, Sonia Seyed Allaei, Steven Vuong, Susan Wang Wang, Tânia Carvalho, Thinam Tamang, Tiago Gonçalves, Tiago Neves, Tomé Albuquerque, Venkata Sowmya Lakshmi Madala, Vinay Varsani, Vitor Esteves, William Guesdon, Xiaoxiao Ma, Yago Bardi, Yassine Baghoussi, Yu Luo, Yuanliu Wanghan, Zeeshan Desai

Advisors: Miguel Castro de Neto and Nuno Santos

WDL Team: Leonid Kholkine, Miguel José Monteiro, Rui Mendes, Fabiana Oliveira, Celso Santana, João Martins, Tammy Contreraz, Margarida Abranches and Shikhar Chauhan

Challenge and Data Providers: Metropolitan Area of Valle de Aburrá, City of Cascais, City of Porto, City of Torino, Nova Cidade Urban Analytics Lab, U-Shift, PSE, CycleAI and OpenWeather.

Financial Support: Fundação Calouste Gulbenkian, ScaleUp Porto, Urban Co-Creation Data Lab, Siemens, basecone, OLX Group, OutSystems, Fidelidade and PSE.

Jury Members: Aarthi Kumar, Alexey Grigorev, Bruno Coelho, Clarisse Magarreiro, Chanukya Patnaik, Diego Esteves, Elisabeth Fernandes, Filipa Castro, Filipa Peleja, Francesco Costigliola, Gilberto Titericz, Inês Teixeira, Jacek Kustra, João Ascensão, João Nunes, João Silva, Julian Miranda, Kelwin Fernandes, Kyra Wullfert, Lisa J. Knoll, Marcel Motta, Miguel Batista, Miguel Cabrera, Nuno Gomes, Nuno Moniz, Nuno Paiva, Paula Alves, Pedro Chaves, Pedro Sarmento, Rajneesh Tiwari, Rita Ribeiro, Sophie Watson and Sudarshan Gopaladesikan

Finals Jury Members: Enrico Gallo, Jan Potter and Steffan Verhulst

Team Mentors: Alina Petukhova, Bernardo Caldas, Boris Tchikoulaev, Bruno Coelho, Carlos Gomes, Carlos Rodrigues, Carolina Ferreira, Clarisse Magarreiro, Daniel Moura, Daniel Rodrigues, Filipa Rodrigues, Gilberto Titericz, Inês Coutinho, Ines Teixeira, Jacek Kustra, João Veiga, Jorge Martinez Rey, Lisa J. Knoll, Kelwin Fernandes, Manuela Almeida, Marcel Motta, Miguel Batista, Nuno Gomes, Nuno Paiva, Pedro Chaves, Pedro Sarmento, Ricardo Vitorino, Rodrigo Coutinho, Sudarshan Gopaladesikan and Tiago Pires

EXECUTIVE SUMMARY

World Data League (WDL) is a data competition that brings data scientists from all over the globe together to solve socially-oriented challenges, focused on the United Nations (UN) Sustainable Development Goals (SDG). In 2021, the first edition of WDL took place with over **14 different challenges** around the topic of Data-Driven Cities, inspired by the 11th UN SDG - Sustainable Cities and Communities.

The main outcome of the competition is open-sourced proof-of-concept algorithms that can help develop sustainable cities. The evaluation process is optimized not only for the technical evaluation but also for understanding the problem, analyses of impact, analyses of the datasets and possible applications of the algorithms in the day-to-day of organizations.

In this document, the authors summarized the insights and findings of the teams for each of the challenges, organized into three main categories:

Approach and Technique describes the technical aspects of the team's submission. A short description of the types of methodologies and algorithms used is presented. The goal of this section is to give an overview on the methodologies that could be used for technical implementation.

Data describes the datasets that the teams worked with, which were provided by real-world institutions. It describes what data the teams found important, where it could be improved and what additional data would have been useful. The goal of this section is to give an idea of what type of data might be needed in order to solve a certain challenge.

Insights and Impact sums up the interesting findings by the teams, either through data analysis or by applying certain mathematical models. The goal of this section is to give an overview of possibilities for insights and impacts that can be achieved with little resources, as the participants only had two weeks to complete the challenges.

To note that the summary presented in this document is exclusively based on the content of the team's submissions and, even though the deliverables have been evaluated by a technical jury, some submissions might be technically inaccurate.

The authors conclude that the first edition of WDL was a success, with many interesting outcomes and models that cities or organizations working with cities can benefit from. In the future, the authors would like to improve the format with more detailed information regarding the team findings. The authors would also like to further support the teams, in order to develop their work. In 2021, WDL had one team that made a scientific publication from the work produced during the competitions - the aim is to increase this number in the future.

GLOSSARY

- **WDL** - World Data League
- **UN** - United Nations
- **SDG** - Sustainable Development Goal
- **PCA** - Principal Component Analysis
- **EDA** - Exploratory Data Analysis
- **ARIMA** - Auto Regressive Integrated Moving Average
- **MSE** - Mean Squared Error
- **MAE** - Mean Average Erro
- **CNN** - Convolutional Neural Network
- **SARIMAX** - Seasonal ARIMA with eXogenous Factors
- **SHAP** - Shapely Additive Explanations
- **LASSO** - Least Absolute Shrinkage and Selection Operator
- **ANN** - Artificial Neural Network
- **LSTM** - Long short-term memory
- **GAM** - Generative Additive Model
- **WHO** - World Health Organisation
- **MLP** - Multilayer Perceptron
- **K-NN** - K-Nearest Neighbours

INDEX

INTRODUCTION

Currently, data is generated in abundant amounts, to such an extent that it is hard for a human on his own to make sense of it all. Luckily, this abundance of data also sparked research into techniques and methodologies to interpret and even predict future outcomes based on this data. And thus, the profession of data scientist was born - a cross between software engineering and mathematical modeling.

These techniques have brought great leverage and advantage to corporations that knew how to use them but being such a recent field, many sectors are still lacking behind in terms of knowledge and know-how. World Data League (WDL) aims to close this gap with organizations that are working on socially-oriented challenges. For this, we connected our challenges to the **United Nations Sustainable Development Goals**.

In 2021, we held our first edition with the theme **Data-Driven Cities** with **over 100 participants** from **34 countries** that worked **14 different challenges during more than 2 months**. It was a very intense and endeavour which produced **over 75 technical reports** responding to the proposed challenges.

How to interpret this document: This document aims at summarizing the used methodologies, showing the conclusions about datasets to solve specific challenges and presenting the main insights found by the teams. The authors would like to stress that the outcomes presented here should be considered as proofs-of-concept with a need for scientific validation. That is due to the fact that participants were limited to the datasets presented to them (which could vary in quality, quantity, and granularity). In many cases, although there is a correlation between certain variables, it should not be considered that it is a direct or indirect cause. The results presented here are a summary of what the teams have presented in their reports. We hope that these ideas can spark future research directions, considerations for the data collected by cities to solve certain challenges and bring new ideas on how data can be leveraged to create social impact. All the ideas presented here can be found in the team's full submissions on the World Data League code repository.

WDL TOPIC: DATA-DRIVEN CITIES

According to the United Nations (UN), 68% of the world population is projected to live in urban areas by 2050. With ever-growing cities, new challenges arise associated with population growth, but also a lot of interesting potential solutions. With the rise of smart city technologies, sensors, and open data initiatives, a data-driven approach is possible to develop those solutions.

That is why we joined in helping to achieve the 11th UN Sustainable Development Goal (SDG) - Sustainable Cities and Communities. All of the challenges and datasets were provided by cities or entities that work towards creating Data-Driven Cities (research institutes, startups, and companies).

The competition was divided into five different topics:

PUBLIC TRANSPORT

TRAFFIC

CYCLING

ENVIRONMENT

NOISE

Outcomes and Insights

STAGE 1

Public Transportation

CHURN MODEL FOR PUBLIC TRANSPORTATION

The public transport system is crucial to support transportation inside a city. However, the system is only optimal if it can serve the population. Network optimization regarding route, stops, interfaces, frequency, and commodities, amongst other issues, is key to achieve this. A common measure to understand the proportion of customers or subscribers who leave a supplier during a given period is called "churn rate". The churn rate is an indicator of customer dissatisfaction when it is high. Studying and even furthermore predicting the churn rate for public transportation can be a good indicator of the quality of service.

Goal: Identify churn profiles and their key driving factors and propose measures to win back lost segments and their expected impact.

Data:

- Demand for public transportation on a semestral basis in each parish of origin and its respective parish of destination in several cities in Portugal. *Provided by PSE.*
- Socio-demographic (age and gender) information of bus users. *Provided by PSE.*

OUTCOMES

Approaches and Techniques

As the data was not very granular, many of the teams focused on data analysis rather than creating predictive models. In most cases, they could already identify the changes in public transportation usage by different demographics. One team proposed to use a Principal Component Analysis (PCA) for finding the main driving factors behind churn. Some teams built predictive models by using either Decision Trees or Gradient Boosting algorithms. One team used K-Means to classify the segments that are churning and later to identify which of the churning profiles affect which route the most.

Data

Besides the data provided, unemployment, parish population data, google mobility, and points of interest (extracted from OpenStreetMaps) were used. Most of the teams agreed that more fine-grained data (ideally daily or even hourly) would be useful towards solving this challenge and the bi-yearly period is too short to make a good prediction. This would also enable the teams to use weather and air quality data. Besides this, it would be helpful to have more segmentation in the social-demographic data, data from ticket validation, car traffic, parking, and mobility data, (from mobile providers). One team suggested using CCTV footage as a way of counting the number of people present at a metro station or bus stop or even inside a bus.

INSIGHTS AND IMPACT

With the limited data available, it was already possible to demonstrate the usefulness of calculating and predicting the churn rate. The teams demonstrated that the demographic distribution of the bus users is different from the population distribution, particularly concerning the younger population. It was shown that the most significant factors for churning were the population density in the district, relative change in unemployment, and age groups.

Many teams pointed out that these outcomes are useful for planning campaigns towards churning groups but also for evaluating current routes. Further improvements could be achieved by improving the quality and quantity of data. With better data, for example, it would be possible to look at the traffic data and calculate the ratio of cost/effort of using a car and public transportation.

The map below is a proof of concept of a solution developed by one of the teams. It shows the variation in usage of public transportation between several locations in the city. The period considered is pre and post-COVID. Although this map is made for a large period, with proper data it can be generated for shorter periods to assist public transportation companies with usage information.

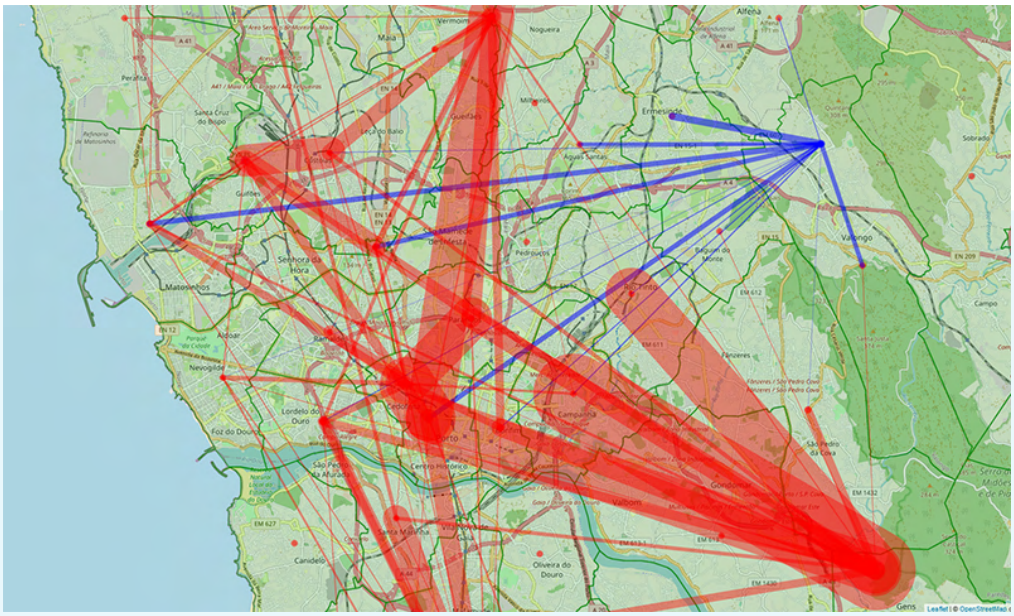


Figure 1

Map representing the connectivity between different nodes in the city - red means a decrease in the usage of the transportation and blue means an increase.

MODEL OF INTEGRATED TRANSPORTS FOR SENIOR CITIZENS

Increasing levels of life expectancy and decreasing levels of fertility are the two leading causes that influence the age structure of the global population, according to the UN World Population Ageing 2020 Highlights.

In recent decades, the elderly population has seen a steady increase as a share of the total population, pointing to an estimate of 727 million people aged 65 or over worldwide.

With increasing age, it becomes harder to perform certain tasks, such as driving, leading to the elderly population increasingly using the public transportation system. With very specific needs and interests, this share of the population is characterized by moving closer to home or within an accessible range by public transport. Simultaneously, they tend to avoid rush hours, and for that reason, a dense offer of stops is crucial.

Goal: Understand mobility patterns of senior citizens, with a focus on providing better conditions of public transport and accessibility to their points of interest.

Data:

- Traffic Intensity Model - the daily average number of senior citizens traveling on road network links between April 2019 and March 2020. *Provided by PSE.*
- Road segments that are part of the different bus routes. *Provided by PSE.*

OUTCOMES

Approaches and Techniques

Since this challenge had a more descriptive goal, all teams focused on doing an extensive data analysis step. This analysis consisted of assessing the correlation between several variables and the average number of senior users of buses for different cities in Portugal.

One team tried to identify clusters in the data, using K-Means and Agglomerative Clustering, with five dependent variables by the district of origin and the average number of senior users, but could not identify any relevant clusters. That same team tried building a predictive model using a Linear Regression, but it yielded a very low accuracy. Another team focused extensively on Graph Network analysis to represent mobility between counties. Their analysis considered the population density of a county, the connectivity between countries in terms of public transportation, and the average usage by senior citizens.

One team approached the problem by trying to identify the best possible location for bus stops using GridSearch considering their distance from points of interest typically associated with the elderly population. A use case was done for healthcare centers, but the team stated that it could be scaled to an even more encompassing set of points of interest, provided that the data was available.

Data

In addition to the provided datasets, more data was used by the teams, such as the purchasing power (which can be related to the cost and quality of life), the criminality rate (which depending on the type of crime can be a deterrent from using public transport), the dependency index of seniors (because the greater the dependency the smaller the ability to use public transport), and weather data (precipitation leads to the non-use of buses). One team used public data about the train routes in order to complement the provided dataset of bus routes. Another team suggested adding to the provided data information about the destination county and the reason why they are using public transport , to identify if the activity being performed could be a key factor - for example, if there was a county where the elderly used public transport to go to the hospital in another county, then increasing travel times or even creating new lines to that hospital could be a good decision. That same team suggested increasing the granularity of the Traffic Intensity Model to a time scale of hours in order to detect the peak hours, and the coordinates of senior-abundant residential zones to increase the number of lines, stops, or both. Increasing the granularity in terms of city would also enable a much more detailed analysis since it was proven that people have different behaviors across cities.

INSIGHTS AND IMPACT

Most teams focused on identifying the variables that influence mobility and the use of buses by the elderly and, in some cases, create a model based on those variables that predicts the number of senior people who use buses daily. Unsurprisingly, the variable that seems to be more influential is the number of links and routes that actually exist - a more robust network, in general, will always lead to more usage. However, in the city of Lisbon, the purchasing power, senior independence, and the number of crimes are also influential.

One team found that, on average, the intra-county mobility is bigger than the inter-county mobility (only one county was an exception to this rule), proving that people tend to move more within their county than to another county. Specifically, in the city of Lisbon, the public transportation network covers extremely well the senior mobility location hotspots, as can be seen in Figure 2. Most of these locations are, in fact, within 400 meters from a form of public transport, with a few exceptions in the mountain of Monsanto. However, not all counties benefit from such a strong network, and one team described in detail the advantages and disadvantages of the public transportation network of each county around the city of Lisbon.

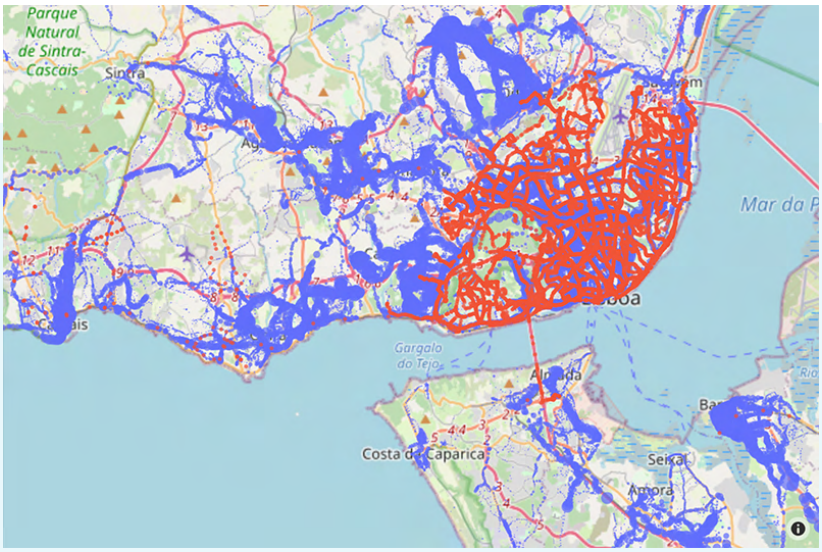


Figure 2
Map showing the senior mobility for the city of Lisbon (in blue) against the bus network offer (in red).

Outcomes and Insights

STAGE 2

Traffic

IDENTIFICATION OF PATTERNS, EXPLANATORY FACTORS, AND PREDICTION OF IRREGULAR PARKING

As the population that lives, works, and visits cities increases, an increasing number of cars circulate on the streets of these cities, especially when public transportation offer is subpar. As a consequence, parking capacity is under significant pressure, with offers seemingly not following demand, which leads to the need for new solutions.

As a starting point, predicting irregular parking can help city administrative services to optimize parking inspection and dissuade possible irregular behavior from drivers. A consequence of this will also be less congestion since there will be minor blockage caused by cars in the streets, which in turn leads to lower CO_2 emissions.

Goal: Develop an explainable predictive model for irregular parking at street level and day.

Data:

- Parking tickets from 2017-2019 in Vancouver, by street and type of infraction. *Open Data by the City of Vancouver.*
- Characteristics of public streets in Vancouver. *Open Data by the City of Vancouver.*

OUTCOMES

Approaches and Techniques

During exploratory data analysis (EDA), teams focused on understanding the characteristics of traffic violations. This included analyzing the location of the infraction, the number of infractions per day, the type of infraction, and even the weather conditions. Analyzing the location of the infraction involved encoding categorical variables. One team analyzed the data in order to determine its stationarity, seasonality, and trend using rolling statistics methods, such as the Dickey-Fuller test.

Due to the nature of the challenge, all teams approached it as a time series forecasting problem. Two teams modelled this problem using an Auto-Regressive Integrated Moving Average (ARIMA) algorithm - one team reported a mean squared error (MSE) of 0.3. One of those teams also used an Exponential Smoothing (ETS) algorithm and a Convolutional Neural Network (CNN), the latter one yielding much better results. Lastly, another team modelled the problem using a Gradient Boosting algorithm.

Data

In addition to the provided datasets, more data was used by two of the teams. One team, which focused extensively on parking infraction causality derived from holidays and other events, gathered a vast amount of data related to bank holidays in Vancouver and major events such as concerts, strikes, and severe weather phenomena during the time comprised by the dataset.

Another team gathered open data about the weather in Vancouver as a way to analyze if it could be a factor or predictor in the number of parking infractions.

INSIGHTS AND IMPACT

During EDA, all teams found a big discrepancy in the number of infractions by type - the following two being highlighted: street infractions and parking meter infractions. All teams also found that, on a road level, there are more parking infractions in arterial roads, such as West Broadway - mainly due to traffic intensity. Teams also found throughout the years, the number of infractions has remained fairly stable.

As mentioned before, one team focused extensively on parking infraction causality derived from holidays and other events and plotted the chart in Figure 3. In red, the days with the least amount of parking violations; in green, the days with the highest amount of parking violations; and in black, the bank holidays and major events in Vancouver. Interestingly, it seems like Christmas Eve (December 24th) is systematically a day with a low number of parking infractions - it was actually marked as red and black in 2017. This could be caused by the smaller traffic flow as people are spending this day in particular mostly at home - however, no such hypothesis was proven using other sources of data. On the other hand, the team did go further into understanding the cause of the three peaks of traffic violations (marked green) by looking at big events on those days. They found that all of them had been days where major events took place in the city, such as concerts.

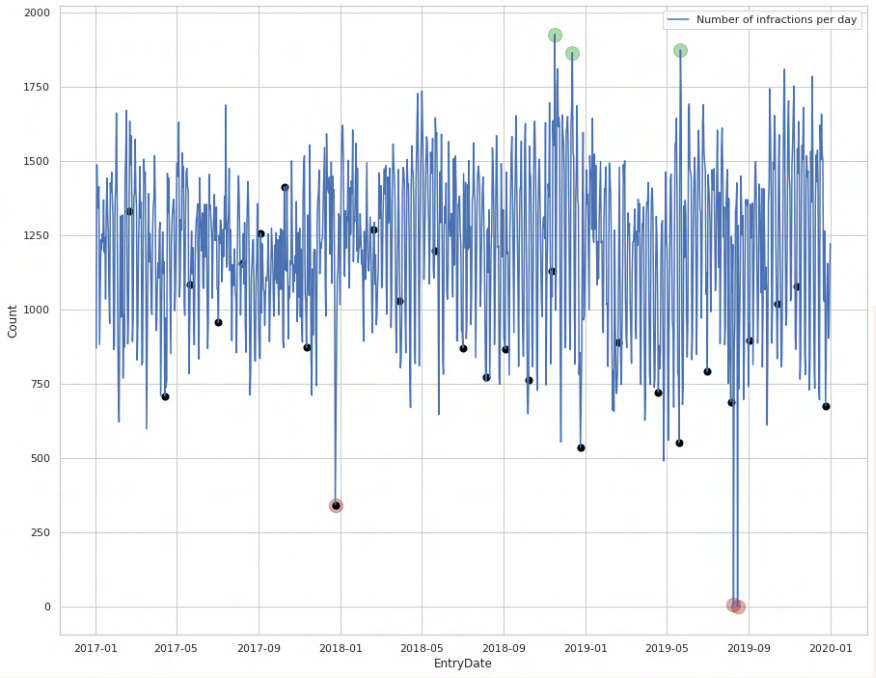


Figure 3

Number of parking infractions (y-axis) per day (x-axis). The red, green, and black dots represent the local minima, maxima and bank holidays, respectively.

Another team drew a map color-coded by the number of parking violations, as seen in Figure 4. They found that northern blocks are where most infractions occur, presumably due to the high concentration of services such as hotels, bars, restaurants, parks, shops, and an airport.

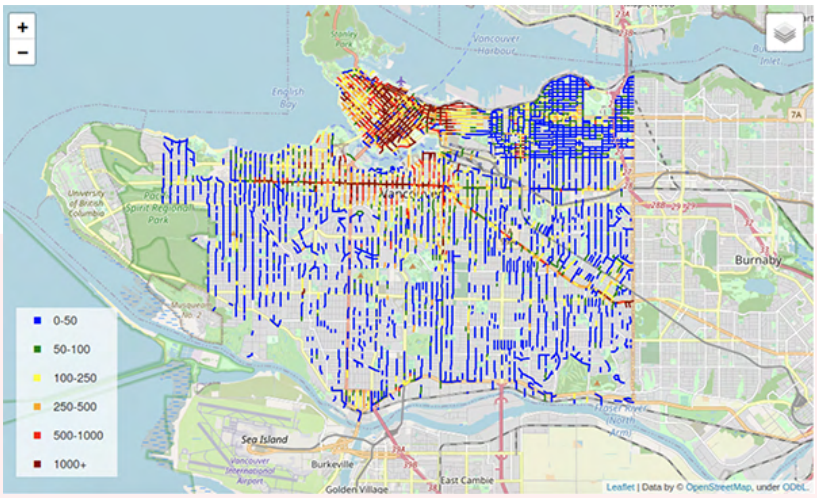


Figure 4

Map of the parking infractions per street in Vancouver.

Lastly, one team also analyzed weather conditions against parking infractions and found that in general precipitation is an impact factor, as can be seen in Figure 5. Days with less rain correspond to days with more parking infractions - which can be expected, as that is usually associated with more people outside on the street and using their cars. When precipitation was above 50%, there was never more than 70 infractions per street daily, for example.

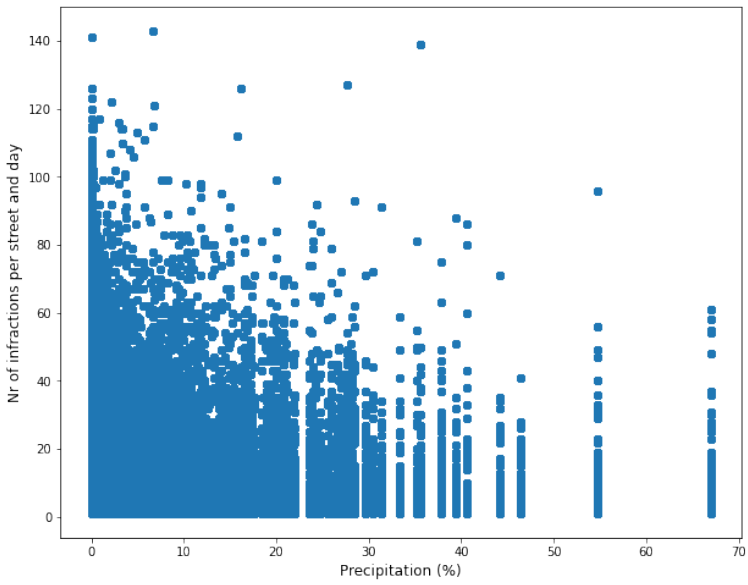


Figure 5

Number of parking infractions (y-axis) compared to precipitation (x-axis).

IDENTIFYING ROAD SEGMENTS WITH POTENTIAL SAFETY HAZARD

According to the 2017 UN's Global Mobility Report, "road transport claims the bulk of transport-related fatalities worldwide: It accounts for 97 percent of the deaths and 93 percent of the costs. On roads, the fatality risk for motorcyclists is 20 times higher than for car occupants, followed by cycling and walking, with seven to nine times higher risk than car travel, respectively. Bus occupants are ten times safer than car occupants. Rail and air are the safest transport modes."

So, decision agents must know where to "improve the safety of mobility across all modes of transport by avoiding fatalities, injuries, and crashes from transport mishaps". That is why creating a model that identifies areas or roads with more considerable risks will help where to invest in road safety and which actions are needed.

The goal is not to predict accidents. Instead, it is to identify hazardous roads so that decision agents take measures on where they need to act in their cities to improve road safety.

Goal: Identify which areas of the city of Lisbon, Porto, or both are riskier by creating a heat-map or intensity model that considers traffic intensity data, characteristics of road segments and their speed profile. This model should identify the most conflictive areas or roads.

Data: A road network level link database with information about its characteristics, including the average speed and the average daily traffic intensity on each segment. It is calculated between April 2019 and March 2020. *Provided by PSE*.

OUTCOMES

Approaches and Techniques

Since no labels were given for what is considered a "hazardous segment", each team used their own assumptions and techniques to classify the road segments. Two teams used the K-Means clustering algorithm to either separate road segments as non-hazardous and hazardous (2 clusters) or give different levels of hazardousness (5 clusters).

As speed limits and speed statistics were included, one team used classification (Logistic Regression and Gradient Boosting classification) to predict if a road segment is likely to have overspeeding and regression techniques (Linear Regression and Gradient Boosting regression) to predict the difference between the average speed and the maximum speed.

Data

Most of the teams found that the data provided was already enough to provide a model for identifying hazardous segments in a city. As external data, one team used OpenStreetMaps to extract amenities and other characteristics found in the road segments, such as police stations, traffic lights, and pedestrian crossings assuming that there could be specific cues that might impact driver behavior. Another team used roundabout presence, rainfall, and geographical data of the parish with its population.

As criticism to the data provided, it was noted that ground truth would be great to ensure that the model itself could be validated. For instance, a dataset of car crashes for the same period as the dataset would have been important. Other interesting data points would be the state of the road and road construction. Some teams also mentioned that provided the data was more granular (e.g., daily basis), it would have been possible to use weather data to enrich the model.

INSIGHTS AND IMPACT

By using data regarding speed limits and the average traveling speed on a road, it was shown by several teams that it is possible to create models which later can be used by policymakers, government officials, and law enforcement to identify which are the most crucial roads for monitoring and intervention. Combining this data with the number of accidents can increase the usefulness of the model. The type of tools that is possible to produce is exemplified in Figure 6, where a map clearly shows which road segments tend to be more hazardous due to an increase in average speed.

One team went a step further and predicted the likelihood of infraction and how much the speed on average deviates from the maximum speed. Their model took into account the presence of traffic lights, pedestrian crossings, and others. Unfortunately, there was no analysis on each effect of the prediction, but that could be performed in the future.

Figure 6
Result of clustering between hazardous and non-hazardous roads - hazardous roads are marked in red.

PATTERNS AND PREDICTIVE MODELING OF TRAFFIC ACCIDENTS

The recent and future increase in population that live and work in cities will significantly pressure the infrastructure of cities, namely roads. This will lead to a rise in the probability of occurrence of traffic accidents, which carries significant challenges in city mobility, transportation systems, and, more importantly, human safety.

In this sense, it is of utmost importance to understand the infrastructural and environmental characteristics of traffic accidents and predict them. This enables, for example, city emergency services to optimize responses to an emergency call and city managers to plan road traffic, considering the risk of traffic accidents.

Goal: Create an explainable predictive model of traffic accidents at street level by moment of the day.

Data: Traffic collision database from the city of Waterloo, Canada, from 2005 to 2018. The dataset included the street of the accident, environmental conditions, and light conditions at the time of the impact. *Open data by the City of Waterloo, Canada.*

OUTCOMES

Approaches and Techniques

In this challenge, a wide variety of methodologies were used for prediction. Some teams that used supervised learning approached this as a regression task (by predicting the number of car accidents on the segment, with a risk factor) or as a classification task (if an accident happened, the categorical target of the number of accidents, location of the accident).
A large array of models was tested by different teams as well. One team compared five models: Random Forest with default hyperparameters and tuning, Logistic Regression, Gradient Boosting with default hyperparameters, and tuning. This team picked Logistic Regression for further prediction analysis as it had higher precision and lower recall. Another team used Random Forest and LASSO, while others used CatBoost and a Neural Network.

One team decided to take an unsupervised approach by clustering areas of accident concentration with DBSCAN. There was also a team that took a time-series approach to predict the number of accidents by day, although they did not develop a street-level model.

Data

Most teams used only the provided dataset. One team also used weather data as input to the model. Teams argued that a more precise model could be built by providing more detailed information (e.g., the hour of each accident and the severity of the accident), information on the quality of the roads, locations of road signs, traffic data, and user behavior (e.g., the demographic of the parties involved in the crash), and data on other means of transportation (e.g., cycling and pedestrians).

INSIGHTS AND IMPACT

One team did an extensive analysis of when most accidents occur and discovered that the number is generally higher during winter and on Fridays (see Figure 7). Most of the accidents happened at or near a private driveway or non-intersection. It was also possible to observe a higher concentration of accidents in other areas.

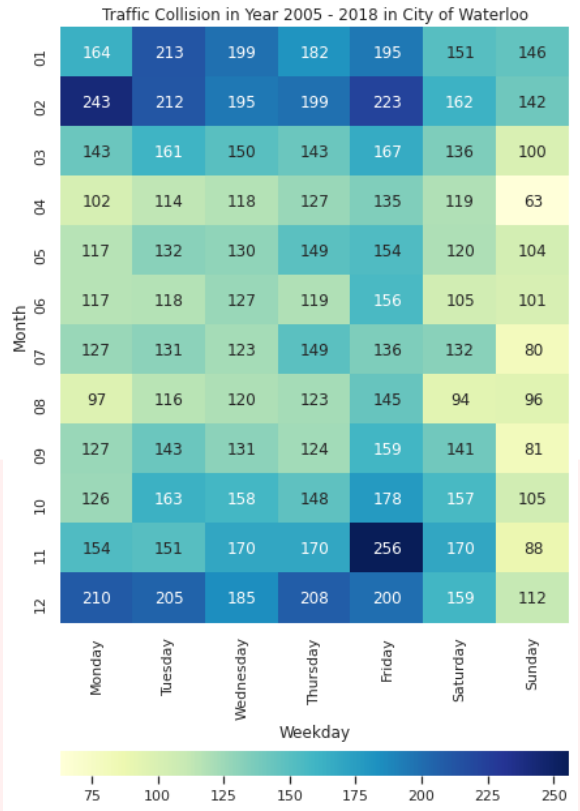


Figure 7

Distribution of the number of collisions between the weekday and the month in the city of Waterloo, Canada.

Another team plotted the accidents, most dangerous areas of the city (see Figure 8), and most dangerous roads. Depending on the features used, the biggest contributing factors were the light conditions, the speed limit of the street and the width of the street. The main opportunity seen for these models is to help local governments and law enforcement to identify dangerous zones in order to improve them or have better patrolling in those areas.

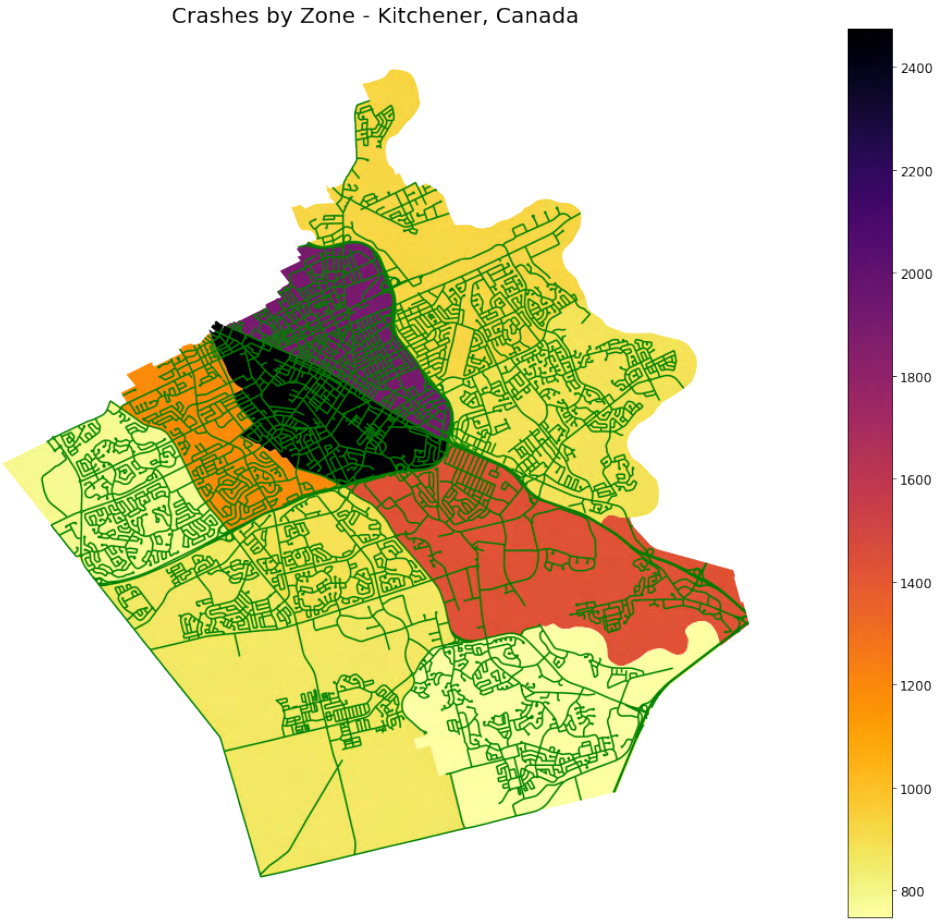


Figure 8

Classification of how dangerous, in terms of car crashes, each area of the city of Kitchener is.

PREDICTING TRAFFIC FLOW IN A CITY USING INDUCTION LOOP SENSORS

Traffic flow in cities is one of the most critical problems to address, as the increasingly higher number of cars circulating in city streets poses a daily problem for city management authorities. In some countries and cities, especially the most populated ones, traffic congestion is such an endemic issue that policymakers and researchers have been trying to solve it for decades and it is still not completely solved.

As a consequence, city navigability becomes severely compromised as drivers have to wait more and more hours in traffic jams. Traffic congestion poses another problem to pedestrians (and to everyone in general) - a negative impact on health. It increases vehicle emissions and degrades ambient air quality, and recent studies have shown excess morbidity and mortality for drivers, commuters and, individuals living near major roadways.

Several cities have been putting forward efforts towards quantifying their traffic levels accurately. The city of Porto, for example, has been counting cars at several locations with inductive loop sensors since 2015. With this historical data, it is possible to predict the traffic flow and especially intense periods of traffic. By predicting traffic flow in the city and understanding the factors that influence traffic, it's possible to take action to reduce it.

Goal: Create a predictive model for traffic flow in the city of Porto for different periods of the day from sensor data and explain which factors impact the traffic flow.

Data:

- Sensor data from induction loops with the number of cars that pass. *Provided by the City of Porto.*
- Location of the sensors. *Provided by the City of Porto.*
- Points of Interests in the city. *Provided by the City of Porto.*

OUTCOMES

Approaches and Techniques

Due to a significant problem of missing values, one team focused exclusively on data from 2019 and added an interpolation step to impute those missing values. That same team also looked for outlier measurements and found very few for that same year, so they replaced the outliers with the average measurement. Both of these adjustments led to a much cleaner dataset. On the other hand, there was another team that focused extensively on outlier removal and data augmentation, which culminated in a quite large dataset for the years of 2018 to 2020.

The first team also looked into autocorrelation and partial-autocorrelation to detect if there was any seasonality in the data, which they found, at seven days. The team tried two modeling approaches - a Random Forest Regressor and SARIMAX - and trained models on a per sensor level, after which they filtered sensors based on the R2-score that was obtained. From 117 sensors, with a minimum R2-score of 0.4, they obtained 71 sensors to work with - the best performing sensor had an R2-score of 0.72.

Another team computed several new features for model training, such as distance to centroid of all sensor locations and many lagged features for the seven days before. They modeled the problem using a Gradient Boosting algorithm, which performed better than a model that always assumes the traffic intensity will be the same as the day before. This team also analyzed the explainability of their algorithm using Shapely Additive Explanations (SHAP), which pointed to the model focusing mostly on lagged features, month features (correlated with seasonality) and some weather features. Another team tried a very similar approach, using a Gradient Boosting algorithm, and obtained a mean error of 0.24, which was much better than the model trained as a baseline.

There was another team that approached the problem as a clustering challenge. They used PCA coupled with K-Means and produced several possibilities using three or four clusters in an effort to understand if different streets could be clustered together in categories. After that, they also did more traditional forecasting methods, even using COVID-19 data and the points of interest.

Data

In addition to the provided datasets, the City of Porto also provided data regarding air quality, noise levels, and weather in their open data portal. The vast majority of teams used several of these additional data sources and merged at least one of them with the traffic intensity dataset.

One team pointed out a major limitation of the provided dataset: the considerably large number of missing values in the measurements, especially for the year 2018.

One team suggested that the small number of air quality sensors compared to the traffic sensors made them quite difficult to use due to the fact that their locations were not the same. They also suggested gathering data related to the position of traffic lights to understand the relation between their functioning and the traffic flow on that position, as well as data regarding the speed of cars that pass on the sensor.

INSIGHTS AND IMPACT

An example of the normal behavior of a traffic intensity sensor can be seen in Figure 9. From 5 am onwards, cars start circulating in that position, and from 6 am to 8 am, there is a sharp rise since that is when people typically leave their homes to go to work. It reaches a peak at around 10 am and remains fairly stable, except for lunchtime, where it fluctuates slightly. At 7 pm it starts to decrease as people are progressively returning home.

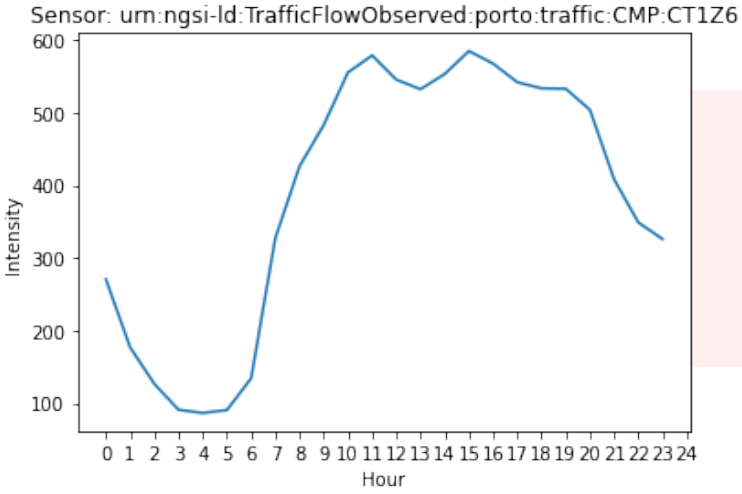


Figure 9
Example of normal behavior for one sensor - hour of the day (x-axis) against traffic intensity (y-axis).

By merging the several sensor data available, one team found that high traffic flow leads to higher air pollution and noise, which matches not only expectations but also literature. They also found that, in general, most traffic flow happens during the weekdays, especially during the peak times when people commute to work and home - 6 am to 8 am and then again 6 pm to 8 pm. During the weekend, traffic intensity is lower in general. Using the model they trained as a source of information, which made predictions for the next 72 hours, this team proposed a solution that revolved around the concept of "Low Emission Zones" - areas of the city where speed limits and car usage is limited. These areas would function with a "color-code system", where there would be three different levels depending on if the traffic flow was low, medium, or high.

Another team analyzed the importance of each factor that contributes to traffic congestion and categorized them into three main groups: weather, time, and points of interest. Interestingly, this team found that the weather has a weight of 18% - this refers to, for example, the impact of rain on traffic congestion, which is a frequent culprit. They also found that time (day, week, month) only accounts for 14%, despite the fluctuation that occurs per time of day and weekday. And lastly, according to this team, the main contributing factor to traffic congestion is points of interest - people tend to move to and from places where there is more infrastructure that attracts them, such as restaurants, shopping malls or parks.

Another team proposed implementing their solution as a smart traffic-light system. By using a forecasting model to understand which zones had higher traffic intensity, a technician could manually tune the behavior of traffic lights in those zones or even have that done automatically. This would lead to decongestion of traffic in these regions and a greater flow of cars, increasing their speed and contributing to the reduction of CO_2 emissions.

One team proposed a traffic monitoring system similar to what's presented in Figure 10. This system would use two colors (red and green), and using the forecasting model would output for each traffic sensor the intensity level compared to the previous day. This way, when the model predicted lower traffic than usual, that sensor would be green, and if the opposite occurred, that sensor would be red. In this case, traffic is mostly at the usual level with the exception of a few places where traffic is expected to be more than usual.

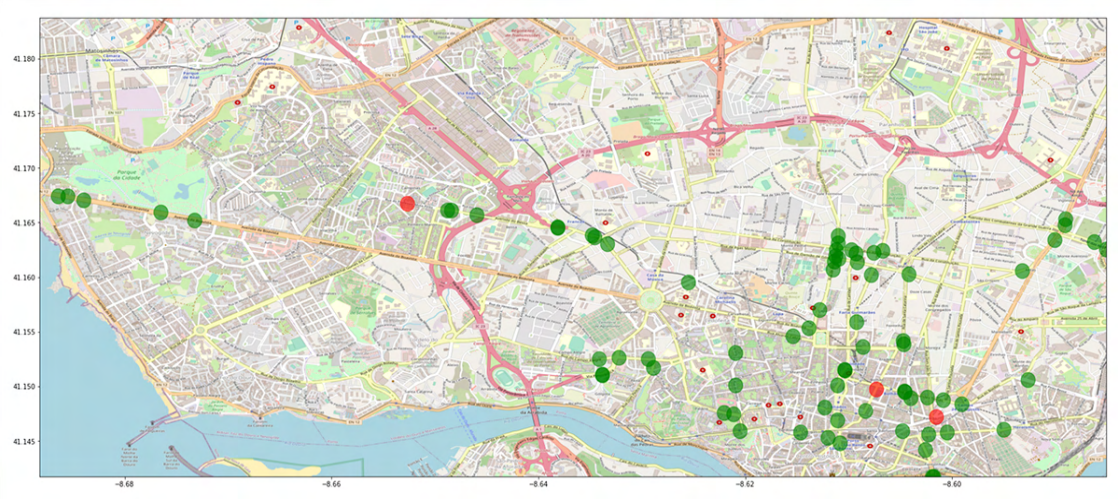


Figure 10

A possible traffic monitoring system, marking when traffic is below usual (green) and above usual (red).

There was one team whose work was published as a scientific paper on ''SoGood2021 - 6th Workshop on Data Science for Social Good'', held in conjunction with ECML PKDD 2021, one of the major conferences in Machine Learning and Knowledge Discovery worldwide. The title of the paper is "Applying Machine Learning for Traffic Forecasting in Porto, Portugal" and can be seen here.

Outcomes and Insights

STAGE 3

Cycling

(LITERALLY) PAVING THE WAY TOWARDS SAFER CITIES

Hundreds of cyclists are involved in accidents on the roads, which results in people hesitating to commute by bicycle because it is perceived as dangerous.

Some of the main risk factors are associated with pavement quality, which is a crucial factor to consider when evaluating cycling safety. Pavement quality refers to the quality of the road when there is no cycle lane or to the cycle lane itself when it exists. The presence of water drainers, potholes, or trail portions contributes to the decrease in pavement quality of cycle lanes.

Decreased quality of a cycle lane can either lead to an accident occurring on the cycle lane itself - for example, a cyclist falling due to a pothole - but also lead to the cyclist not using the cycle lane out of safety concerns, which forces the usage of the road and increases the risk of accidents with motorized vehicles.

Goal: Create a high-resolution map of Lisbon with an embedded layer of pavement quality.

Data: Dataset of 10 000 pictures of roads in Lisbon. *Open Data by Google Streetview API.*

OUTCOMES

Approaches and Techniques

Since this problem was a Computer Vision challenge and the provided dataset was not annotated, all teams had to define an alternative solution to train a prediction model. Additionally, teams defined "pavement quality" differently, and for that reason, they focused on different scopes of the issue.

One team framed the problem as one of pothole detection. They started by segmenting the images on the dataset to identify only road segments and exclude the rest. For that, they used a pre-trained Keras model that was trained on the Cityscapes dataset, which is available as open data. Afterward, they manually annotated part of the dataset using HyperLabel to train a YOLOv5 algorithm. They trained this model in 3000 epochs, following the recommendations from a Roboflow notebook. After detection, the team then computed the area of the road segment and the area of the pothole for each image and by calculating the ratio between the two they used as a measurement of the risk rate of each image.

Several other teams used the same algorithm but framed the problem differently. For example, one team used an external dataset that was already annotated, the RoadDamageDetector, which not only saved the effort of manually annotating data but also enabled them to identify which types of pavement defects to focus on, since it comes with preloaded classes. This team also used the Pareto Theory to identify which pavement defects to focus on first since the distribution of the occurrences of these defects is not equal. On the other hand, there was a team that, while also using both YOLOv5 and a pre-annotated dataset for training, resorted to a different dataset - the Kaggle Pothole dataset. As the name indicates, this dataset focused exclusively on potholes, which limited the scope of the solution.

On a different note, there was a team that approached the problem in a very different way from a technical point of view. This team also did manual annotation of images but established their own annotation policy with the intent of detecting what they perceived were risk factors in pavement quality. They looked at three characteristics, each one with its own classes - street width (single car, double car), pavement type (tar, cobblestone, unpaved), and pavement quality (low, high) - and assigned a risk factor to each class in order to differentiate higher and lower risk. The team then built a car detector using YOLOv5 and assumed that the presence of detected cars in a certain region could be indicative of the risk level for cycling. They built three different predictive models - one for each category - and then averaged out the predictions along with the assigned risk factor.

Data

Mostly due to the fact that the provided data did not contain annotation, teams had the option of either manually annotating data or gathering pre-annotated open datasets. Most teams resorted to the second option, which led to teams using either the RoadDamageDetector dataset or the Kaggle Pothole dataset.

The first one was produced regarding a data competition in 2020 and contains data from Japan, India, and Czech Republic. This data consists of pre-annotated images with classes that represent several types of pavement defects, such as cracks, crosswalks and line blurs, and potholes. The second dataset was produced during a Kaggle competition and contains 665 images of roads with the potholes labeled.

Another team gathered open data about the cycle network of Lisbon as a way to relate it with the provided dataset. This team suggested adding more data regarding cycling accidents, time-series data with the number of bicycles crossing a certain area, and a pre-segmentation of the road section for model classification.

INSIGHTS AND IMPACT

In the end, all teams managed to produce a map of Lisbon indicating the pavement quality on the locations where the dataset images were taken. An example can be seen in Figure 11. Most teams agreed that a similar model could be applied to any city, which makes it very scalable. Teams also suggested that a tool like this would be useful for city planners to know where to focus in terms of repairs so that cyclists would feel safer.

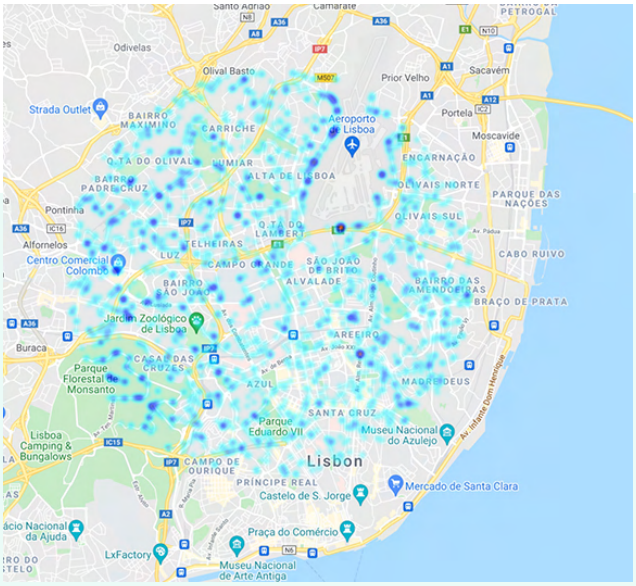


Figure 11

Map of Lisbon with a color-coded indication of the pavement quality (low pavement quality in darker tones and high pavement quality in lighter tones).

One team did a data analysis on their predictions and found that the vast majority of images had no defects, but when they exist, the most common are potholes. However, the team does not suggest starting by fixing all the potholes. Instead, they propose a more sophisticated strategy for tackling the pavement quality issue, following their Pareto distribution analysis: starting by fixing the 20% of roads that have more than 50% of the defects. This team also produced one visualization map per defect and a data-driven strategy for pavement maintenance strategy based on Uber's H3 approach.

Another team also analyzed their predictions and found some key insights on their data regarding streets in Lisbon - for example, the fact that one single avenue had 29 potholes and seemed in urgent need of repair work, especially considering its proximity to the airport. They also analyzed in this specific avenue how much time would be saved upon fixing the potholes.

Lastly, building on their proposed solution, one team made suggestions for both software developers and policymakers. For the second stakeholder, they concluded that although their algorithm outputs the risk level on a map, there are several high-risk zones that are not appropriate to have bike lanes at all. Hence, this tool could work as a decision support system with a visual interface to help decision-makers.

MISSING LINKS - CLOSING THE CIRCUIT IN EXISTING CYCLE NETWORKS

In many places, particularly low cycling maturity cities, several investments in cycling infrastructure are made without considering the overall network impact. One of the most frequent examples is placing a piece of network segment that does not connect to other cycling network parts, despite research showing that a cycle network retrieves more modal shift gains if well connected.

The result of subsequent such practices is a cycling network that is made of several segments that do not join together anywhere instead of a proper well-connected network, which results in users not being able to use the bicycle fully as a means of transportation. Planning of the infrastructure needed to change mobility patterns requires high-quality evidence. Data and models are needed to ensure cost-effective investment, enabling new infrastructure to be planned where most needed. In the case of cycling, it has the greatest potential to replace unnecessary and unhealthy short motorized trips in more dense urban centers, which means designing efficient cycleways and safe street connections.

Goal: Create a data-driven framework/solution for finding the missing segments in a scattered cycling network of a city.

Data:

- Lisbon Cycling Network, containing the traces and description of all the cycling paths in Lisbon. *Open Data by CicloviasPT*.
- Origin-Destination Matrix with the number of trips by modes of transportation in Lisbon. *Open Data by INE*.
- Digital Elevation Model of Lisbon. *Open Data by U-Shift*.

OUTCOMES

Approaches and Techniques

Due to the nature of this challenge, all teams focused extensively on data analysis and exploration. Some techniques used were merging different data sources, data normalization, and plotting information in map visualizations.

One team computed distance centroids per district of Lisbon to identify locations where there is a lack of bike lanes close to those district centroids. They handled data from 24 Lisbon districts provided with 3502 bike lanes. In addition, they analyzed the mobility patterns of each district to find relevant discrepancies based on data from other means of transportation, such as car, metro, and train. They also interpreted the problem as a Graph Network analysis issue and therefore built a graph network to analyze the inter-connectivity between districts, from which they derived several conclusions.

Data

Besides the provided data, the data provider suggested other data sources, such as the UK National Travel Survey and the Sydney Travel Survey and cycling count data, which has information about the cycling habits of people in the UK and Sydney.

Teams also used open data from GIRA, the shared bicycle system in place in Lisbon, and information about the location of public transportation such as train and metro, in which a bicycle could be transported.

INSIGHTS AND IMPACT

One team used data from several means of transportation to understand where citizens usually want to go as a way to identify places where cycle networks should exist. On top of that, they used external data from the shared bicycle system of Lisbon, which, coupled with bike usage data, enabled them to understand where new bike segments should go. With that in mind, they identified the districts with the highest mobility (from all modes of transportation) and highest relative bike usage. Using that information, the team identified the districts with the most unnecessary car trips (because of their shortness) that could easily be done by bike instead.

Figure 12

Map of the current cycling network (black lines), shared bike system stations (blue), public transportation (red and black), and inclination of streets, from heavy to flat.

Lastly, this team also produced the visualization in Figure 12, which gives an overall idea of the cycling network in Lisbon and how connected it is with public transportation. It also shows the inclination of roads, which was an additional factor that this team considered when suggesting new segments of bike pathways.

According to this team, the main missing link is connecting the current cycle network that ends at Avenida da Liberdade with Terreiro do Paço through Baixa-Chiado. This segment is crucial, considering the inclination of the road (completely flat), the presence of bike sharing stations, the connectivity to public transportation (bus, metro and ferry) and the mobility patterns of citizens. Additionally, in their work, they detail more suggested segments and the specific districts and areas of Lisbon.

Another team focused on connectivity between districts by analyzing the number of cycling roads that connect each one of them. Some districts are very well connected, both intra-district and inter-district, and others lack inter or intra connectivity. For example, the districts of Ajuda and Campo de Ourique are the only two districts completely disconnected via bike from their neighbors and, for that reason, they could be good candidates for connection.

This team also suggested that a good criterion for connecting two cycling lanes would be to consider the shortest distance, meaning that current bike lanes that are closest to each other should be connected. Alternatively, by calculating the centroid of mobility of each district, another criterion could be the shortest distance to that centroid.

PREDICTING THE DEMAND FOR SHARED BICYCLES

Bicycle sharing systems are becoming increasingly popular in cities, especially the ones with a flat topography. The systems operate by means of stations where users pick up and drop off bicycles as they use them. These stations can be manual or automatic.

In manual stations, a person is responsible for lending and receiving the bicycles to each user. In contrast, in the automatic stations, there are a limited number of anchor points from which the bikes can be loaned and to which you have to hook the bike that you want to return. If users wish to return their bike and do not have empty anchor points, they will have to move to another station or wait for another user to release the anchor point.

One of the main challenges of these bicycle sharing systems is the lack of predictability of usage to keep constant availability of bicycles and anchor points. In other words, the system has to keep bikes available by the time the user approaches the station to pick up a bicycle and, on the other hand, have anchor points available for the moment the user wants to drop it off.

It can be considered to a certain extent that the flow of passengers is usually from the same origin to the same destination. For example, in the mornings, university students often travel from subway stations to make the last mile to their institutions by bike.

Goal: Create a model that indicates the optimal movement of bicycles to be made between stations and at what times or frequencies - load balancing system.

Data:

- Information about the bike loans, from 2014 to 2021. *Provided by the Metropolitan Area of the Aburrá Valley, Colombia.*
- Location of the stations. *Provided by the Metropolitan Area of the Aburrá Valley Colombia.*
- Digital Terrain Model, containing the elevation of the city. *Provided by the Metropolitan Area of the Aburrá Valley, Colombia.*
- Location of public transportation stations, namely bus and metro. *Provided by the Metropolitan Area of the Aburrá Valley, Colombia.*

OUTCOMES

Approaches and Techniques

All teams focused extensively on exploratory data analysis in order to understand the mobility patterns of the bike-sharing system. The first steps on that exploration were spent cleaning data, after teams realized that the dataset had some inherent errors - one team did a particularly thorough data cleaning process. This same team then focused on the bike station that had the most observations (pick-ups and drop-offs) and analyzed it even further. They looked at the correlation between weather and the difference between loans and drop-offs and, after that, they performed autocorrelation on an hourly and daily level. Regarding modeling, they firstly focused on predicting the difference between pick-ups and drop-offs for that specific station, using a Random Forest Regressor on a particular period - obtained results were an R2-score of 0.47. They then extended this model to a group of bike stations, namely the top-15 bike stations, for which the R2-score ranged from 0.00 to 0.86.

Another team focused on characterizing stations according to their different patterns of usage and found three different types: stations that receive more than they lend, stations that lend more than they receive, and stations that receive and lend in the same proportion. For that, they mainly resorted to data visualization. This team used a CNN to predict the maximum balance between pick-ups and drop-offs for each station on each day and obtained a Root MSE of 9.34. They then used the predictions of this model to build a graph that performs load balancing between nodes. This graph was built using heuristics, considering, among other factors, the distance between stations.

After EDA, one team approached the problem as a forecasting task - for the bike demand prediction - followed by spatial data analytics - for exploring route options between bike stations - and lastly, an optimization step - for defining how many bikes should be transported. Due to lack of time, this team only produced a model for the forecasting task. The team used a Gradient Boosting algorithm for a forecast of one day and two days in advance, with a Root MSE of 17.2 and 18.6, respectively.

Data

Several teams noticed errors in the provided dataset, such as duplicate entries, missing values, drop-off times before the pick-up times, and loan durations that were too long (e.g., more than one year).

One team used weather data as a possible correlating factor with bike usage, and another gathered data of bank holidays in Colombia to analyze its impact on the number of bike loans. There was also a team using a combination of both.

INSIGHTS AND IMPACT

Several teams noted the fact that the median duration of a bike loan is 14 minutes.

After an extensive exploration step, one team found several insights regarding the historical usage of the bike-sharing system.

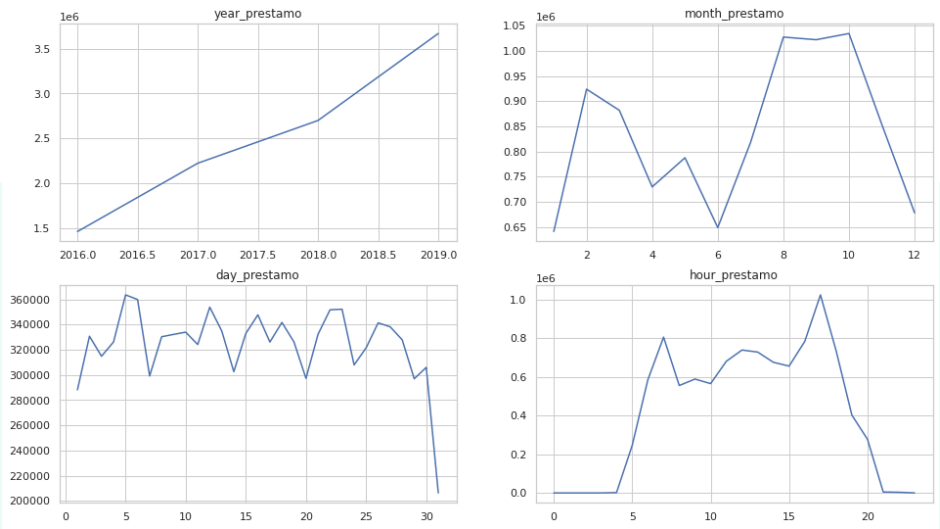


Figure 13

Several charts representing the historical usage of the bike-sharing system, comparing the number of loans (y-axis) with different time scales (x-axis) - top left: number of loans throughout the years; top right: average number of loans per month of the year; bottom left: average number of loans per day of the month; bottom right: average number of loans per hour of the day.

We can see that, over the years, there has been a constant increase in the use of shared bicycles, which proves that more and more people are using them.

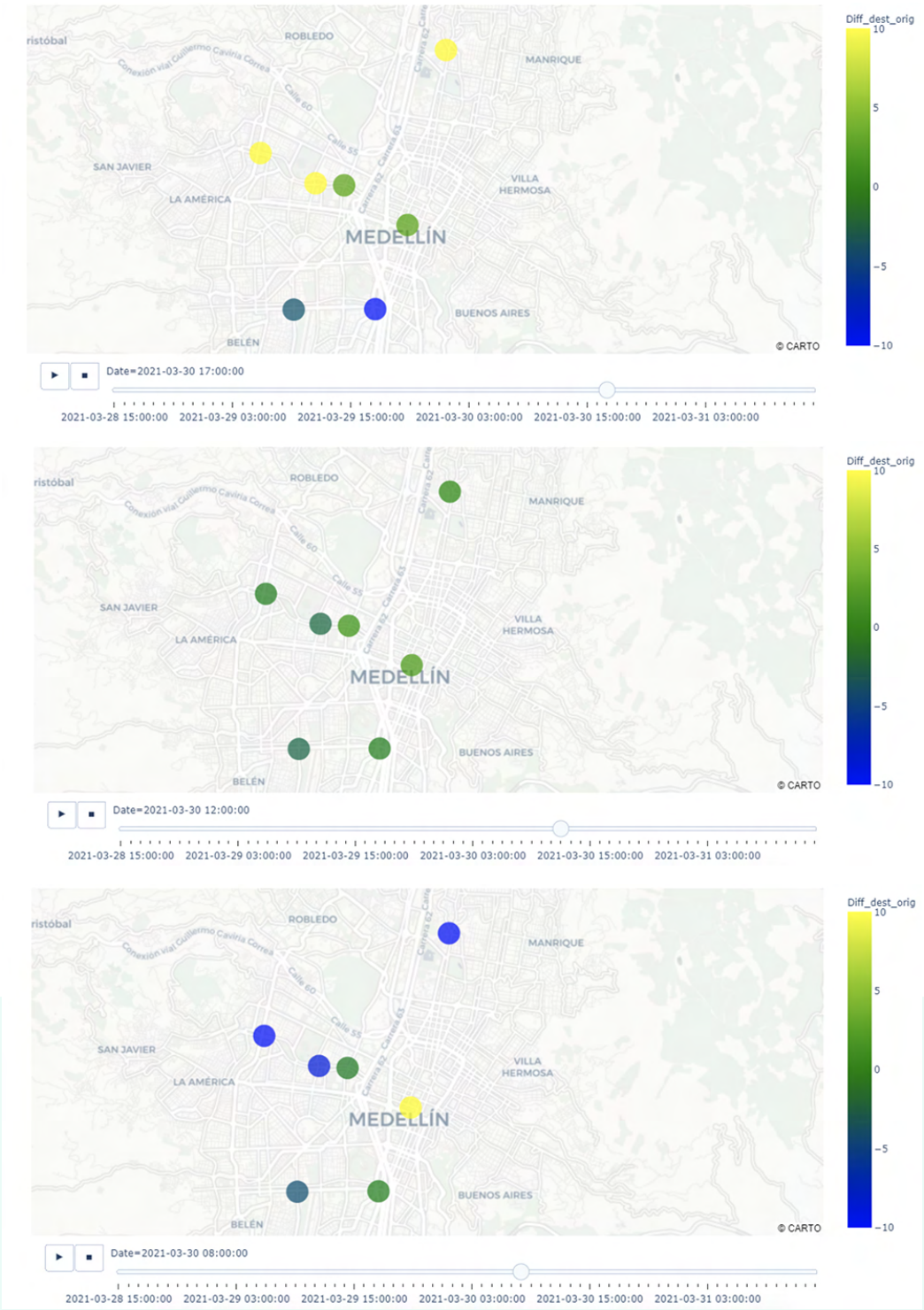
Regarding the monthly usage, August, September, and October are the months with more usage, while January, June, and December are the ones with the least. This could mostly be due to weather conditions and the number of tourists.

Over a single month, the number of loans per day remains quite similar, with minor fluctuations.

Throughout the day, there are two peaks that stand out and that coincide with the rush hours (one in the early morning and the other in the late afternoon). There is also a peak at lunchtime. These three moments of the day are the busiest because they include commuting home-work, work-home, and even traveling to lunch. It is also possible to identify approximately the opening and closing times of the stations - 4 am and 10 pm, respectively.

This same team also focused on characterizing stations according to their different patterns of usage and found three different types: stations that receive more bikes than they give, stations that give more than they receive, and stations that receive and give in the same proportion. They found that different types of stations also have very different types of usage behavior. For example, one station had a usage trend that has been consistently rising throughout the years. In contrast, another had a trend that had virtually stayed the same throughout the years and that had a big drop in the month of December.

Another team, whose technical approach focused on the bike stations with the highest amount of loans, noticed that the top-15 bike stations account for 45% of the bike demand of the whole system. This team also proposed a product solution based on their technical developments. They suggested a Bike Balancing Map that used a color-code system to represent stations and were balanced, stations that should receive bikes, and stations that should give away bikes. They tested this solution on six bike stations and showed how depending on the time of day, certain stations should receive bicycles (blue) from other stations (yellow) and that some stations are balanced (green).



[Figure 14]

Bike Balancing Map at three different times of day - morning (top), afternoon (middle), and evening (bottom).

Outcomes and Insights

STAGE 4

Environment

ATTRACTING POPULATION TO GREEN SPACES IN METROPOLITAN AREAS

According to a UN report, "urban green areas offer great opportunities for positive change and the sustainable development of our cities" due to creating spaces for outdoor activities. Also, according to the World Health Organisation (WHO), "urban green spaces, such as parks, playgrounds, and residential greenery, can promote mental and physical health, and reduce morbidity and mortality in urban residents by providing psychological relaxation and stress alleviation, stimulating social cohesion, supporting physical activity, and reducing exposure to air pollutants, noise and excessive heat.".

The goal of the challenge is to understand how socio-demographic factors, tourist attractions, and people's mobility level around the green spaces explain the demand for these spaces.

Goal: Create a model that predicts the average daily demand for green spaces and the main contributing factors.

Data:

- Visits to green spaces, extracted from mobility data. The data was a snapshot in time and not time series. *Provided by PSE.*
- The number of museums, parking lots, buildings, families, and people residing in the surrounding area of the green space. *Provided by PSE.*
- Percentage of residents that are younger than 19 years old or part of the senior population. *Provided by PSE.*

OUTCOMES

Approaches and Techniques

Since the dataset was quite small, the teams started by examining the correlations between the feature sets. Afterward, the teams tried to train different regression models with the target set as the demand. One team tried linear regression, LASSO regression, ridge regression, and Random Forest algorithms, with very large MSE and MAE (Mean Average Error). Another team tried even a bigger range of algorithms: Linear Regression, Decision Tree and Random Forest Regressor, K Nearest Neighbor, Ridge Regression, Bayesian Regression, Principal Component Regression, Polynomial Regression, and Partial Least Squares Regression. The teams reported very weak results mainly due to the size of the dataset. Most of the teams used the feature importance to analyze the possible factors that were most influential.

Data

Three major criticisms to the dataset were identified: lack of clarity on how the area of influence is calculated, the size of the dataset and the imbalance between the number of green spaces in the city of Porto and Lisbon. The participants noted that it would be better to work with raw data and, ideally, use an automatic visitor counting in the green areas to yield better results.

One team supplemented the dataset with OpenStreetMap data (from cafes and restaurants, for example), parish socio-demographic information, and pollution levels from OpenWeather. Other teams have also suggested adding extra data, such as the location and data regarding safety in the area of the green space, as well as the time that is spent in the green area.

INSIGHTS AND IMPACT

The teams found it hard to compare the data between Lisbon and Porto, as Porto has five times fewer parks than Lisbon. Nevertheless, it was noticed that what holds true for one city, might not be for another - for example, in Lisbon, the elderly population has a positive correlation with the demand, while in Porto, it is negative.

Regarding the main driving factors, it was discovered that accessibility to the park (see Figure 15), playgrounds, and more pedestrian streets are paramount. The predictive power of the models was meager, as the dataset was very small and without any temporal information.

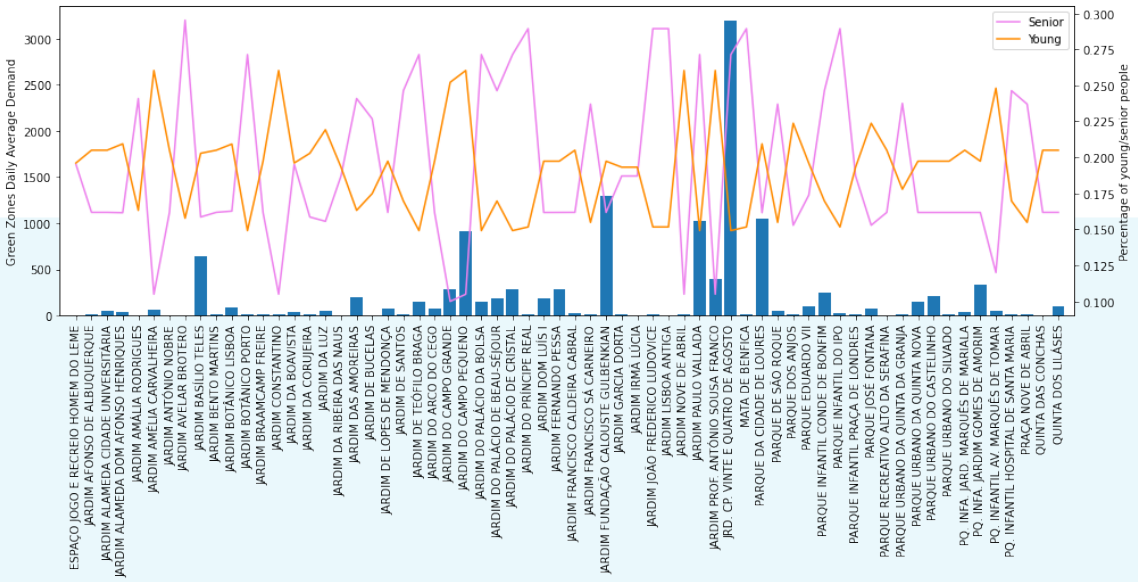


Figure 15
Demand of the different green spaces in Porto - the orange line represents the younger population while the purple one represents seniors. The green space with the biggest demand is Campo 24 de Agosto, which is well connected to the different transport systems.

PREDICTING AIR QUALITY FOR OUTDOOR ACTIVITIES

According to the WHO, there are several health consequences of poor air quality, such as an increased risk of respiratory infections, heart disease, and lung cancer. While reducing air pollution is a slow process, small improvements can already impact the health of many citizens. This challenge focuses on predicting the ideal locations for outdoor activities, taking into account the air quality in the city of Cascais.

Since 2020, Cascais has been implementing an amount of air quality sensors. This air quality system aims to monitor what's happening in Cascais and identify areas where there's a need to act and improve air quality while improving and creating better experiences for its inhabitants.

Goal: Create a model that predicts the best locations for outdoor activities by minimizing the effect of air pollution.

Data:

- Air quality measurements from 11 sensors with daily averages. *Data provided by the Municipality of Cascais.*
- Points of interest for outdoor activities. *Data provided by the Municipality of Cascais.*

OUTCOMES

Approaches and Techniques

One team created a Multilayer Perceptron (MLP) to classify the air quality into five different levels and then used K-Nearest Neighbours (K-NN) to predict to which outdoor activity location each sensor belongs.

Another team did an EDA to find initial patterns in the data. Afterward, the stationarity and autocorrelation of one of the stations was analyzed by applying the Dickey-Fuller method. Finally, the future values were predicted by applying ARIMA.

Data

One team used the weather information from OpenWeatherMap to complete the information missing from what was provided by the city of Cascais. They also added extra information regarding traffic on the roads in a 3km radius from the sensor. It was suggested to include the UV measurements in the future as it might help predict ozone levels.

INSIGHTS AND IMPACT

With simply an initial analysis, it was already possible to find where in the city the air quality is not as good - for example, Guincho had several indicators above the median of the city in terms of Nitrogen Dioxide, PM10, and PM2.5. Another analysis showed how much time each parish spends under different levels of pollutants.

The teams that participated in this challenge presented a tool that could be used in the future to show locations with outdoor activities and the risk of air pollution (see Figure 16). Predicting the pollution levels in the future could also be helpful in regulating traffic and reduce the air quality health impact or to create targeted marketing campaigns to raise awareness among the population.

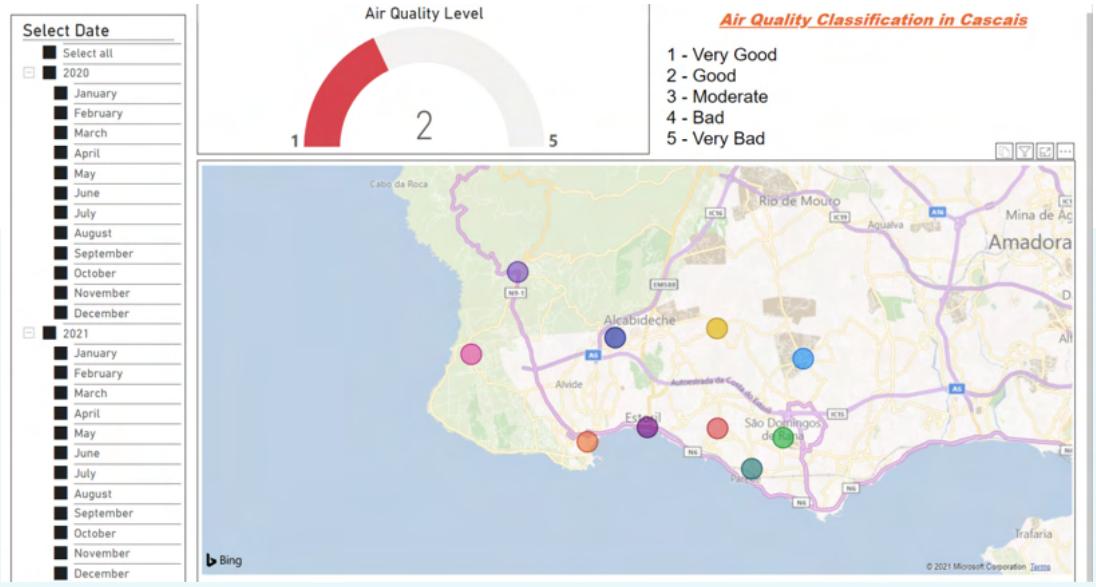


Figure 16
An example of a dashboard showing the predictions of possible locations for outdoor activities.

OPTIMIZATION OF THE OUTDOOR ADVERTISING WITHIN CITIES

This challenge tackled a less discussed problem in cities: visual pollution. Many cities are flooded by countless outdoor advertising panels. It is known that visual aspects are crucial in the urban planning process since each modification can generate obstruction to urban elements. Besides the city aesthetics, it is also possible to argue that too much outdoor advertising can decrease wellbeing and increase social inequality. It is clear that reducing outdoor advertisement can create a positive impact on the city. The goal of this challenge is to do so with a minimum impact on the reduction of the audience.

Goal: Create a model that optimizes the number and location of the outdoor advertising positions in order to minimize the visual impact in urban environments. A better dimensioning and integration of outdoor advertising positions in cities should also be considered.

Data: List of outdoor advertisements in the entire territory of mainland Portugal with the location, number of average views, its maximum visibility, and height. *Provided by PSE.*

OUTCOMES

Approaches and Techniques

Due to the nature of this challenge, very different approaches were proposed by the competing teams. One proposal was to use a clustering algorithm (K-Means and K-ProtoTypes) to cluster the location of advertisement placement and to remove the advertisements with fewer views from the cluster.

Other teams decided on a metaheuristics approach. One approach was based on a local or neighborhood search that optimizes for the spread of advertisement panels while at the same time maintaining the number of views high. Another approach was to train a regressor with a Gradient Boosting model to predict the number of views at a particular location, generate random geographical points, and for those points predict the number of views. Later on ,optimize on a set of parameters the points to be selected (e.g., spread from concentration points or number of views).

One team proposed three methodologies: removing panels with less contribution, spreading out panels by selecting areas with the biggest viewership, and a genetic algorithm.

Data

By using domain knowledge, it was noted that some panels were missing in the dataset. It was also recommended to include in the dataset the type of outdoor advertisement, advertisement cost, and the direction in which the advertisement panel is facing.

More concretely on the dataset, it would also be essential to have the time frame to which the dataset is related to (e.g. does it include the COVID period?) and how are the different metrics calculated (e.g. daily views and max visibility).

Other interesting data to be considered in the future can be the socio-demographic information of each of the parishes and the companies that run the billboards to minimize the problem by company.

INSIGHTS AND IMPACT

As expected, the biggest urban centers have the highest concentration of outdoor advertisement (see Figure 17). But also the distribution of the viewership was not equal. https://bit.ly/3qyemLY 56.7% of panels produced 80.6% of total views, and 71.4% of panels produced 90.4% of total views. This shows the potential for a better distribution of the outdoor advertisement. It was noted that if it were possible to predict the number of unique viewership for each geographical point (e.g., by using mobility data), it would be possible to create an optimization model which could be used for optimizing the distribution of outdoor panels while also minimizing the number of panels.

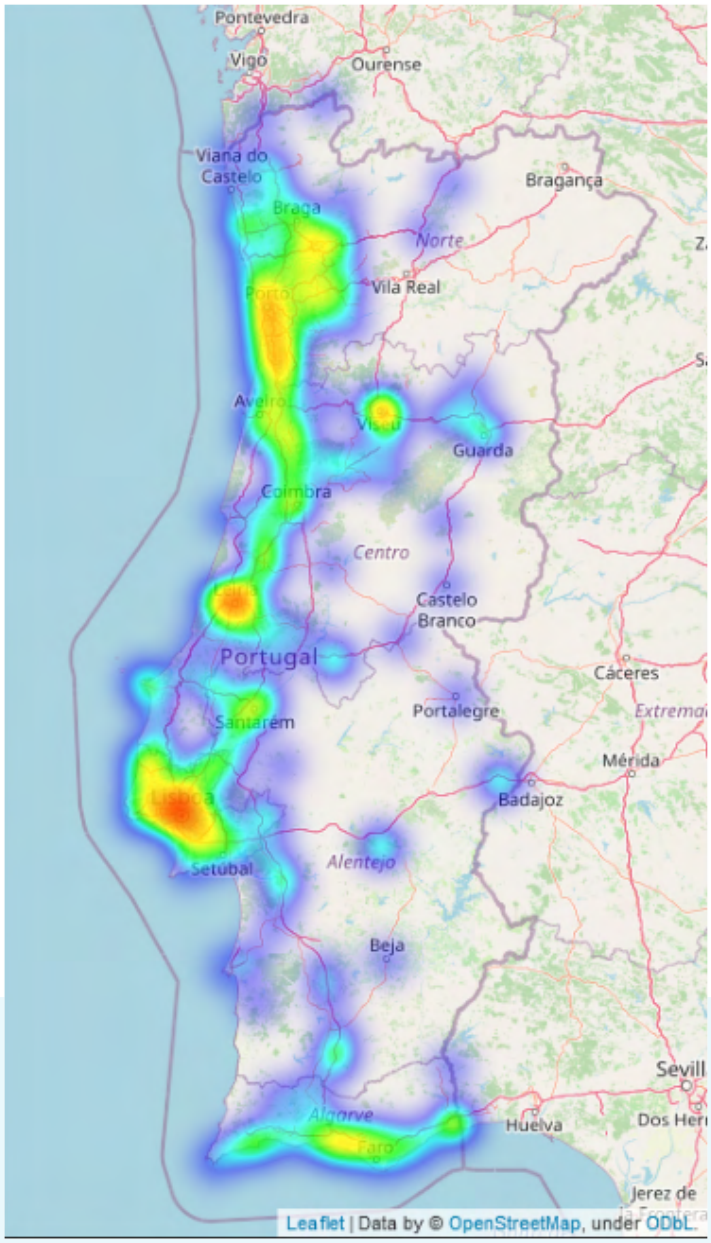
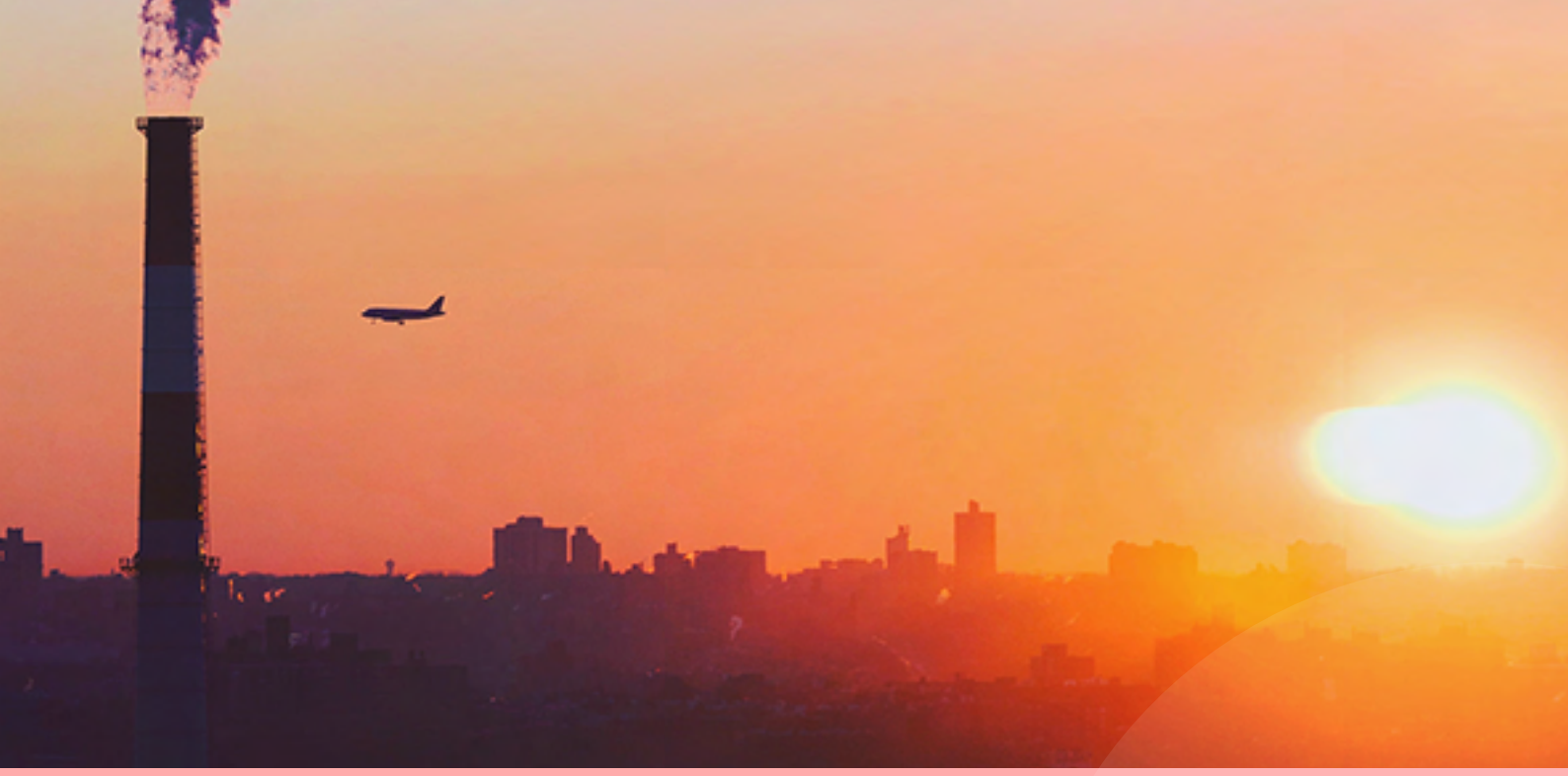


Figure 17
Distribution of outdoor panels in continental Portugal - most of the panels are present in Leiria, Porto, and Lisbon.



Outcomes and Insights

FINALS

Noise Pollution

IMPROVING THE QUALITY OF LIFE BY REDUCING CITY NOISE LEVELS

According to the European Environment Agency (EEA) report on environmental noise in Europe, health risks are posed when the population is exposed to increased noise. Some vulnerable groups have been specifically identified:

- In children, exposure to aircraft noise can affect cognition skills in school;
- The elderly are more vulnerable to sleep disturbances, and noise during the night can affect their rest and have a negative impact on cardiovascular diseases;
- Pregnant women are also more vulnerable to sleep disturbance, and environmental noise may also increase the risk for preterm and low weight birth;
- Socio-economically disadvantaged people might also be at higher exposure to noise levels due to poor housing conditions, pre-existing health conditions, or fewer opportunities to cope with noise.

However, it is not only humans affected by noise, but the report also shows that biodiversity negatively affects terrestrial and aquatic species.

This shows that there are several benefits to reducing noise pollution in cities. One source of noise pollution is the recreational nightlife noise, which comes from loud conversations during nighttime on the streets.

The city of Torino has been studying the noise levels of recreational nightlife noise in the San Salvario area, which is home to many bars and clubs. They have installed several sensors in order to measure the noise levels with records since 2016.

Goal:

- Predict future noise levels and, if possible, explain complaint trends that can be attributed to leisure noise levels.
- Build a model that could predict noise in recreational nightlife, especially peaks of noise outside what is considered normal.
- Study the feasibility of predicting the complaints related to noise levels.
- Suggest a framework of how these models can be integrated into the city's decision-making process and allocation of resources.

Data:

- Population by census micro-areas. *Provided by the City of Torino.*
- Pubs, restaurants, and other businesses in the San Salvario area. *Provided by the City of Torino.*
- The noise level measurements from IoT sensors in the San Salvario area. *Provided by City of Torino and ARPA Piemonte.*
- Municipal police complaints to the police, including related to noise and leisure noise. *Data provided by the City of Torino.*
- Number of people in different locations of the San Salvario area extracted from mobile phone users and aggregated by age segments. *Provided by Olivetti.*
- Number of people in different locations of the San Salvario area extracted from WiFi users. *Provided by H2020 Rock Project.*
- Other georeferenced data was available from Torino's open data portal.

OUTCOMES

Approaches and Techniques

Most of the teams conducted an extensive EDA to understand the data by looking at daily averages and how the different variables were correlated. One team noted that it is vital to use logarithmic averages for calculating averages during a certain period. However, there were different modeling approaches for predicting noise levels: some used tree based algorithms (such as XGBoost, Random Forest and linear regression for baseline model), others used time series (SARIMAX and Moving Average) and Artificial Neural Network (ANN) such as CNN and LSTM (Long Short-Term Memory).

One team experimented with GAM (Generalised Additive Models) to associate to sound intensity the features. They have also developed a bivariate multimodal distribution from gaussian kernels. Another team focused on predicting outliers in data to detect unexpected changes in noise. Regarding the feasibility study for predicting complaints, while some teams did a theoretical analysis, others developed models by applying tree-based algorithms.

One team used a K-NN algorithm to obtain patterns characterizing demographics in order to derive those profiles with a higher propensity to complaining.

Data

The amount of data used by the teams also varied. Some teams decided to use less datasets to develop the solution. Other teams incorporated data points such as weather from Open-StreetMaps, holidays in Torino, academic calendar, COVID restrictions, football matches in Torino, and business locations. One team encoded the COVID measures into six different categories and did other feature extraction such as converting the night before a holiday, Friday and Saturday in holiday days.

The teams had two criticisms towards the dataset: a larger dataset regarding the number of people at a specific location would have been useful for a better estimation of how the number of people can influence the noise levels. The other criticism is regarding the police complaints: Many complaints have missing hours of complaint.

Other interesting data to include when modeling mentioned in the submitted solutions included socio-economic data, local events, business opening hours, more precise location of complaints, a better context of the complaints (e.g., party next door, party on the street), and more sensors in the streets.

INSIGHTS AND IMPACT

During the initial analyses, it was found that the obvious assumption holds true: COVID (see Figure 18), weekends and holidays, weather, and football games influence the noise levels. A surprising discovery was that most of the complaints happened during the week, when the noise level was lower compared to the weekend. It was also found that many of the complaints are logged in the morning during low noise levels. While some teams attempted to predict complaints, the results varied. The general conclusion is that it is not possible to predict complaints based on noise levels, as the current data needs more quality. One team proposed to create an app where Torino's residents could notify the police anonymously and could also collect the data about complaints. Another team suggested creating an annoyance score to enrich the complaints dataset by combining the probability of noise exceeding the threshold level, causing annoyance and causing a complaint.

The teams saw that these models could be used to predict noise levels and apply preventive measures during more problematic situations. One team recommended creating a traffic light system to inform the authorities about the upcoming noise level changes. Another team noted that it is essential for the model to be explainable, as it can give a better intuition to the authorities of where intervention might be needed.

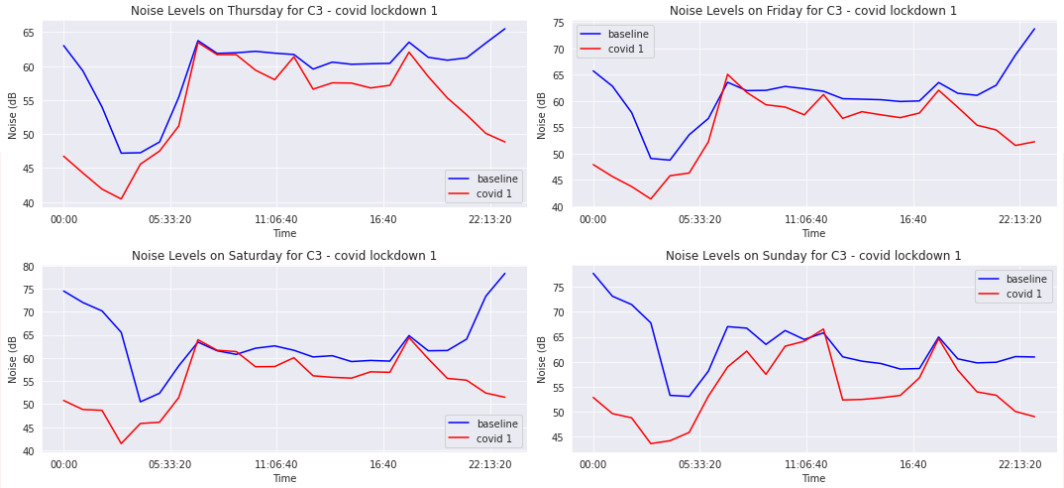


Figure 18
Variation of average noise levels for one of the sensors before the COVID lockdown (blue line) and during the COVID lockdown (red line). While, on average, noise levels during the day remain similar, they are much lower at night.

CONCLUSIONS

In this report, a summary of the work done by more than 100 participants over the course of several months is provided. As an outcome of the competition, 75 technical reports were gathered on 14 different challenges. All the technical reports include code, which we hope can be used to prototype and develop the solutions further in the future. It is important to mention that these reports differ in terms of quality, which can be assessed by the position of the respective team on the leaderboard.

One of the teams has successfully submitted a scientific paper to SoGood 2021, a workshop of the ECML-PKDD 2021 conference.

The challenges provided to the participants were also very varied in terms of field of study - optimization, time series forecasting, computer vision, geospatial analytics, among others - and also the size of the datasets and the type of available data. On top of it all, the data was provided by real-world institutions or from open data portals, which posed a true test and challenge to all participants in terms of data-centric development.

As a future direction, WDL would like to make the reports from each challenge even more accessible to a non-technical audience. We would also like to support teams that want to develop their ideas further and ensure that the methodologies presented are scientifically rigorous.

We hope that these results can spark future research directions, and provide feedback to cities on how data can be leveraged to solve their challenges.

INTERESTED IN OUR WORK OR WOULD LIKE TO SUBMIT A CHALLENGE?

Get in touch with us at
hello@worlddataleague.com

FOLLOW US ON

Linkedin

Twitter