Title on Deep Stuff

Project report

Alaeddine Yacoub 03675497 | Kheireddine Achour 03730007 | Mohamed Mezghanni 03730431 | Oumaima Zneidi 03675592

Salem Sfaxi 03729393 | Stephan Rappensperger 03735303 | Yosra Bahri 03666016

September 4, 2020

Motivation

The COVID-19 pandemic is a great challenge for most countries. In order to fight the spread of the pandemic, Multiple measures were taken ranging from partial to complete lockdown and the shutdown of all businesses that were deemed unnecessary. After a while, the impacts of these measures on societies and economies began to manifest. Starting from June, governments began executing their exit strategies in order to begin recovering from the societal and economic impacts and reopen to the outside world. With the summer holiday nearing, people get to go to some of their favourite holiday destinations. As social distancing is still recommended in most cities, it is expected that various touristic destinations would lose their attractiveness while other may feel an increase in popularity.

The management of the tourism flow of Munich is central to any policy concerning the COVID-19 crisis since tourism accounts for a significant share of the GDP and economic growth of this region.

Over the course of this project, we set to make use of these results in order to build a tourism clustering and recommendation system for the post-COVID period in the city of Munich.

---

1 Project Description

The aim of this project is to tackle the change in several aspects of the tourism flow due to the COVID19-crisis in the city of Munich. Using a wide spectrum of data-driven approaches, we are intending to build a model which will allow us to analyze the tourism flow in the next period and deliver a touristic recommendation system for Munich post-COVID19.

This system will use key factors and different indices of Munich' tourism venues helping us cluster the attractions and unveil the tourism patterns trends and changes amid the corona crisis. We will also provide insights of various tourist hot spots that gained in popularity after the crisis. This system will deliver a holistic picture of tourist movements and preferred spots to businesses and regional decision-makers which will ultimately lead to policies that optimizes the region's tourism sector.

2 Research Question

How did the tourism patterns change over the course of the COVID-19 crisis and how could we create a reliable touristic guide system for the city of Munich while exploring these patterns pre-/post-COVID19 and the impact of infection cases on the dynamics of Munich city?

3 Goals

Our main objective in this part is to contribute to optimizing tourism flows throughout the city its venues and attractions. Over the course of this project we will deploy various data-driven approaches to understand and classify classes of tourism attractions according to their appeal to different groups of visitors and the targeted audience. This will then provide insights of tourism markets segmentation in Munich and evidence of various tourist patterns to consider when presenting policies concerning the COVID-19 crisis.

4 Data basis

Harvesting key factors such as the number of visits for each venue, reviews, ratings, satisfaction levels, visit period, origin of the visitor and trip type (solo, couple, family, business and friends) of Munich' tourist attractions and turning them into a database, helped us cluster these attractions and unveil the tourism patterns changes amid the corona crisis.

Another aspect of the data basis was the number of the COVID-19 cases registered in Munich. These changes and fluctuations were used as input for our different models in order to select an appropriate venue according to the local pandemic situation.

The datasets were used differently for each part of the project. However, all datasets were merged for the recommendation system.

5 Sources

For the datasets collection we used multiple sources:

- TripAdvisor
- Airbnb
- Bavarian State
- Robert Koch Institute (RKI)

6 Data collection

TripAdvisor

As previously discussed in the second milestone, the web-scraping approach was used to extract datasets from " TripAdvisor. The focus was brought on the top ten attractions in Munich and their visitors. For each tourist attraction (e.g. Marienplatz, Allianz Arena ..), an updated dataset is available presenting several features about the visitors. The features used from the TripAdvisor datasets follow this scheme :

- Date: month and year of the visit of the tourist to the attraction.
- Visitor origin: city and country of origin of the visitor
- Type of visit: this feature describes the type of travel (family, couple, friends...)
- Rating: rating of the touristic monument out of 50

These datasets are used to analyze the origin of the visitors in Munich and its evolution pre- and post-COVID-19. The visitors are first grouped by their country of origin. Afterwards they are grouped by the month and season to get a better idea of the evolution of the travel activity. We decided to only focus on the time period between August 2019 and August 2020.

Airbnb

As it is more adequate to use the Airbnb data directly from the website, all available datasets of the listings from February 2019 until June 2020 were downloaded and pre-processed in the python file Airbnb data.py for further use in the data processing pipeline. The difference between those datasets is that each month the values of all features would be continuously updated. Theses features contain details about:

- neighbourhood: the city district
- id: the accommodation ID
- number of reviews: The number of reviews given by the users to each accommodation
- price: The price of each accommodation per night
- last review: The date of last review concerning the accommodation

This information will help us to analyze the change of the Airbnb accommodations pre- and post-COVID-19. The accommodations were grouped by their neighbourhoods, which allowed us to have the number of reviews (e.g number of tourists who stayed at these places) and prices in each city district of Munich.

Bavarian State

The file munichdata.py is used to download the visitor figures of the most important public places in Munich from January 1, 2000 as an Excel file. For better processing of the data, the visitor figures are summarized in the munichdata.csv file. After pre-processing the data, the file consists of several columns, each column representing a location. A small selection of places is: Olympia Park, Bavarian National Museum, Bavarian State Orchestra, German Museum, etc. Each line of the file represents one month. Since this file also uses the date as index, this record can be easily extended. For example, for Hotspot Prediction the number of visitors is combined with the number of corona cases.

6.1 Robert Koch Institute (RKI)

We use the Robert Koch Institute (RKI) as our source for the Coronavirus data. Using this script rkidata.py we first download the current case numbers via the RKI API. To be able to process the data better afterwards, the .json file is converted into a Pandas DataFrame in the next step. Since this data is only used in conjunction with other data sets, it is very important that the individual files can be easily combined. For this reason, the date column is set as the index of the file and converted into the form "day/month/year". Since we only need the trend of corona infections for the Munich district for the machine learning models, all additional information are removed from the file. As a last step this script removes all special characters (e.g. umlauts) and saves the edited file as a .csv table in the folder "data/covid19data/...".

Pre-processing

6.2 Tourist Flow and Clustering Models

To cluster the attractions of the city we used the information retrieved from TripAdvisor. We grouped this based on the group type of the visitors (alone, as a couple, with friends, with family or on a business trip). We also retrieved the origin of each visitor (if the visitor is local, from within Bavaria, from Germany, EU country or foreigner). This analysis involves the following three pre-processing steps.

- Scraping, Loading, and processing the datasets: Individual datasets for each attraction from the top 25 attractions in the Munich's region needed to be retrieved and features need to be extracted and aligned. In order to do this, we created the data processing pipeline tourism flow data to be found in the src/ folder. This contains a feature extraction() and get visitors() function in order to calculate the number of visitors, group them by type and create a dataset ready for clustering
- Handling Categorical Variables and Scaling the data: After extracting all the necessary numbers of visitors by type and origin and handling the categorical variables (using One-Hot Encoding) we needed to scale these since our model works with an Euclidean distance measure and is therefore sensitive to magnitudes.
- Split Ratio and applying the clustering model: To cluster our data we applied a simple K-Means Clustering Algorithm. The output of this analysis is a set of medoids which are the representative for each class of attraction based on its visitors and a vector of clustering labels. We used all the data to create the clusters.

For the trajectory analysis, we are only interested in the country of origin of the visitors of a certain attraction. Therefore, we will be extraction only that information and dropping all remaining columns. Then, we will group all entries from the same country in order to get the number of tourists from that country that have visited the attraction. The origin of a user usually follows this simple scheme "City, Region, Country". However, not all three locations are always given. If the location does not contain the country or an incorrect area (such as Yugoslavia), then that row will be removed from the dataset. Finally, we compute the percentage of incoming visitors from the same country to the attraction.

6.3 Hotspot Forecast Model

In order to predict how crowded the respective public places in Munich are, several individual data sets are combined with each other. Thus, the data from TripAdvisor, Airbnb and government serve as a basis for the visitor numbers. The RKI, on the other hand, provides the data on corona infections, as these can have a major impact on the number of visitors. In order to generate a data set for model training, the following steps are performed:

- Scraping and downloading the data: Separate scripts have been programmed for downloading the respective data sets. These scripts can be executed separately and save the respective data as a csv file.

- Formatting of the data: To combine the individual data tables into one gigantic table, the format, i.e. the date and the places, is standardized. In the Airbnb and Munich dataset only the format of the date has to be adapted. Since the web scraper we programmed for the TripAdvisor data creates a separate file for each public place in Munich, we first read all places into a data frame. Then we adjust the places and date format like did beforehand with the two otherone. In case of the RKI dataset the numbers of one month are summed up. The time before corona will be filled um with zeros.
- Merge and save: Now the four datasets - TripAdvisor, Airbnb, Munich visitor numbers and RKI - are being merged to one table and saved as a csv file.

6.4 Recommendation System

As the recommendation system will use all of the available places in the datasets pre-processed for the Tourist Flow and Clustering Models as well as the extracted Munich data from Bavarian State, both of Munich visitors and TripAdvisor csv files were used in this step.

For the TripAdvisor data the column date and visit was cleaned by eliminating the extra expressions such as 'Date of experience' and 'Trip type'. The date format in the date column was also changed and sorted in the descending form to be used as an index later on. The column origin was split into three separate columns 'city', 'country' and 'extra'. A further column of the provenance type was added by distinguishing between the cities and countries the visitors came from (either inside or outside EU, inside or outside Munich etc..).

For the city portal datasets unwanted columns such as "Ausland(Tourismus)" and "Inland(Tourismus)" were dropped and the column Date was replaced by a Year column for the purpose of grouping all values yearly and summing them to get the total visitors count That will be playing a role in weighing the importance of each place. One of the most criteria a place will be evaluated with is the already predicted metrics for the touristic hotspots pre- and post-COVID19, for this purpose, an 'all score metric' column will be also extracted from the predicted forecast dataset depending on the date of visit of every user.

Other columns were added for these datasets to determine the city district each place belongs to using the geo-coordinates saved in the "geoattractions.csv" file, as well as the type of place (indoor or outdoor) by splitting all the touristic attractions into two lists according to their type.

7 Data Models

In this section the structure of the different machine learning models for the individual project parts will be explained in detail. To simplify the implementation of the models the Python library Keras with TensorFlow as back-end was used. With the help of this library the models easily can be composed of single layers.

7.1 Clustering model

During the previous milestone, we worked on a the understanding and classification of classes of tourism attractions according to their appeal to different groups of visitors and the targeted audience. To unveil the tourism markets segmentation and patterns, we implemented and compared the results of two clustering algorithms: K-means clustering and DBSCAN.

Approach

In our previous deliverable, we used the information retrieved from TripAdvisor to cluster the top 25 touristic attractions from Munich. We grouped this based on the group type of the visitors (alone, as a couple, with friends, with family or on a business trip). We also retrieved the origin of each visitor, this information is also given by the user and usually follows the format "city, region, country".

During this milestone, after selecting the most appropriate model, we went through the following steps.

- Tune the model's Hyperparameters
- Applying hierarchical clustering instead of simple k-means and comparing the results
- Visualizing the clusters on a map

Training

Being an unsupervised machine learning clustering algorithm, we trained the k-Means with the whole data points as input to groups them into 4 clusters. This process of grouping is the training phase of the learning algorithm. The result would be a model that takes a data sample as input and returns the cluster that the new data point belongs to, according the training that the model went through. The number of clusters was determined by the elbow method which is a heuristic method consisting of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.

Evaluation

Comparing the results from the DBSCAN and the K-means yields a better result in favor of the K-means for the reason that we have a sparse data for each attraction (not all the visits type or origin were recorded for each review). This was shown by the Area Under the Curve (AUC) that can be found in the example notebook.

As for the the tourist flow analysis, we still can notice limitations to our approach concerning the accuracy. First of all, we are only relying on the Tripadvisor dataset. This means that we have no information on visitors that are not using Tripadvisor. German locals can also visit Munich and would not require Tripadvisor since friends or family will recommend locations to them. In turn, they would not leave a review and will therefore not be considered in our model.

7.2 Hotspot Forecast

As already described in the last milestone, this section of the program aims to predict the number of visitors to various public places in Munich on a monthly basis. The goal is to determine the amount of visitors for the next four months as accurately as possible.

To do so, a machine learning structure is needed that can process time series. So-called recurrent neural networks (RNNs) are among the most suitable networks for such tasks. These models are able to store information over several time steps. The long-term short-term memory network (LSTM) is a special type of RNN and probably the most common one.

In addition to the LSTM network, a one-dimensional neural network will be analyzed and optimized. Subsequently, the two models are evaluated and compared based on their performance.

As data basis, we combined the data from TripAdvisor, Airbnb, the government of Munich and the Coronavirus infections of the RKI. The resulting file has the structure:

- Date: month and year of information
- Visitor numbers: The number of visitors to the various public places from the government file. Using these numbers, we created an importance metric that classifies the different tourist sites according to their popularity.

Approach

Since the visitor numbers of the different public places in Munich were listed as a time series, a method must first be written that determines input and output. Therefore a method was implemented that uses a defined number of months for input and stores the following four months as output. Then the starting month is incremented by one and the process is repeated. This is done until the end of the time series is reached. Now the input data is stored as a three-dimensional array of the form [samples, time steps, features]. The output has the form [samples, forecast time steps]. With this approach it is important to note that the number of months used for an input sample is now also an optimizable parameter. Also we have to store for each place a separate model.

The next step is the scaling of the rearranged dataset. For this, the standard scaler of the library Scikit-Learn was used. To be able to determine the performance of the machine learning models

later on, the data set is divided into training and test data. The ratio is 2/3 training data and 1/3 test data. This division of the data is a frequently used one.

Training

In training step, the two artificial neural networks are now optimized. First, the general structure of the network is adapted to the input data. Also the number of months used to generate an input sample is optimized.

Then the individual hyper parameters, such as the size of the one dimensional CNN core and the number of neurons in a layer, are being optimized.

To speed up the training of the models, the early stopping of the library Keras was used. This function checks after each epoch whether the mean squared error of the validation data is still reduced. If it remains higher for 10 epochs, the training is stopped and the best value of the trained values is being used.

Evaluation

Now we use the test data set to perform the evaluation. The performance of the models can be determined in both ways, graphically and analytically. For the graphical evaluation the correlation between prediction and real value will be plotted. In the optimal case the resulting points lead to a bisection of the first quadrant. However, this method is only suitable to get a rough overview.

For a more precise evaluation, the correlation of the four predicted months is calculated for each public square.This should be in the range between 0.8 and 1. In addition to the correlation, the mean squared error is determined for each place. The lower this value is, the more accurate is the prediction.

To simplify the evaluation, mlflow is used. With mlflow the different model parameters as well as the correlation plots etc. can be stored. The models can be easily compared with each other.

7.3 Recommendation System

The touristic hotspots in Munich pre- and post-COVID19 have diverse criteria that could match the preferences of the visitors differently. To distinguish between those hotspots according to the personal choice of each user, it was important to build a recommendation system that is based on the predicted (importance) metrics for each touristic place and the gathered features from the datasets of the official city portal and TripAdvisor

For our last objective, we aimed to work on recommending some of the most popular local venues in Munich to a specific user.

Approach

Recommendation systems are usually executed with two different approaches : the collaborative based filtering and the content based filtering, the latter is mainly relying on recommending an item to a user which is similar to previous items highly rated by this same user. On the other hand the collaborative filtering is more based on the user behaviour, which means that people with similar characteristics have a higher chance to share a similar taste and that exactly fits to our expectations in building this recommendation system. Hence we decided to go further with the collaborative filtering approach.

Training

First Method : k-Means Clustering After extracting the TripAdvisor data and pre-processing it, a K-Means model was used for clustering the visitors according to multiple features that depend on two main features : visitor origin and visit type. The elbow method helped us to determine the number of clusters (10 clusters) in the data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The visitor origin will be divided into four cases :

- The visitor came from a country outside the European Union : provenance Outside EU
- The visitor came from a country inside the European Union apart from Germany : provenance EU apart from GER
- The visitor came from Germany apart from Munich : provenance outside Munich
- The visitor came from Munich : provenance Munich

The visit type will be divided into five cases :

- The visitor came with friends : visit Traveled with friends
- The visitor came with a partner : visit Traveled as a couple
- The visitor came with family : visit Traveled with family
- The visitor came alone : visit Traveled solo
- The visitor came on business : visit Traveled on business

After the clustering process and basing on which cluster the user now belongs to, a score prediction function will take in consideration the preferences of the other visitors to predict the score for each place and how similar it is comparing it with the user expectations.

Second Method: Scoring System Another method that contributed to the outcome of the recommendation system is a scoring system function, which is different from the above mentioned score prediction function (used only with the places extracted from the TripAdvisor datasets). Since many places could only be found in the city portal, it was necessary to adapt the recommendation system using this second method. After saving the preferences through the user interface, a scoring function will be assigning a score to each place basing on the similarities between the places

features and the entries, for example, if the user prefers indoor activities then all places, which are of type indoors will get 10 points to their score, if the user is staying in a specific district then all places which are located in that district will also get extra points. These points will add up to each place and give the first weight used in the scoring system. The second step is to also use weights such as number of visits and most importantly the predicted metric (importance) of each place as a touristic Hotspot post-COVID19 after giving suitable coefficients to each weight.

The user entries will include this list:

- visitor origin: This parameter will check on the percentage of visitors coming from the same origin to each place using the already described origin analysis in the subsection **??**.
- visitor accommodation: The district in which the user will be staying (in case visiting for more than one night)
- activity type: if the visitor prefers to have activities indoor / outdoor
- visit type: Type of visitor (business, single, couple, family, group)
- date of visit : This parameter will define the month of the predicted metrics (weights of the hotspots)

Evaluation

This combination of both methods k-Means clustering and scoring system allowed us to include diverse criteria and provide the user with a more flexible and efficient recommendation system, that would not only depend on the available database of TripAdvisor and the city portal before and after the covid-19, but will also emphasize the preferences of each user in the aim of providing the most adequate listing of the touristic hotspots. To have an accurate evaluation of the implemented recommendation system we had used several combinations of user entries and observed the results basing on each preference. This method of evaluation helped us to ameliorate the functionality of the code and accuracy of results.

8 Results

In this chapter, the analytical results of the different methods will be presented separately. After presenting the results of the used methods, a discussion over the outcome and further optimization possibilities will be conducted. By the means of different scenarios, the strength points and limitations of each model will be explained in detail. At the end we will showcase the trends highlighted by our models.

8.1 Clustering Models

Observations

After running the clustering Algorithm, We can proceed with the visualisation. For that we will be using folium to create a map with the help of the Leaflet library. We first start by assigning the geocoordinates of each country to a point using the geopy library. Lastly, we create a marker on the map on those geocoordinates. The size of the marker of country is linearly dependant from the number of visitors originating from it. As for the clusters, we used a color coding to highlight diffrent clusters and their respective attractions. All of the graphics can be found on our website.

Trends

In the flow analysis model, we can recognize a clear trend: The visits after the relaxation of the lockdown measures are more concentrated in the EU region and internal tourism. This is a logical consequence of the travel ban/warning handed over to the other non EU countries. Before the pandemic, we can clearly visualize the origin diversity of the incoming visitors. As for the clustering a clear trend change could not be showcased.

8.2 Hotspot Forecast

As already mentioned, two models (CNN and LSTM) have been optimized for the hotspot forecasting system. The performance of both models is determined by calculating the parameters correlation and mean square error using the test data. Since a separate model - but always with the same general structure - was trained for each public place, this is now also taken into account in the evaluation. For a better overview the determined values are listed in tabular form. Tables 1 and 2 show the values of the optimized one-dimensional Convolutional Neural Network. Tables 3 and 4 show the values of the long-term short-term memory network.

Observations

So-called correlation plots were created for the graphical evaluation. As an example, the figure 1 shows the correlation between the predicted numbers for the second month and the true visitor numbers of the Olympia tower. Each of the blue data points in this diagram represents a sample from the test data set.

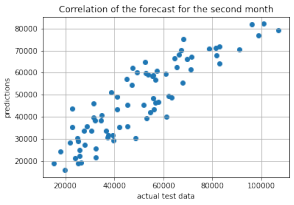

Figure 1: Correlation plot of real value and prediction from the Olympia tower

Trends

During the training of the two models different trends could be identified. At first it can be observed that the correlation is lower for public places with a high number of zeros or no visitors in the data set than for places that had visitors at any time. In contrast, the correlation decreased slightly with increasing distance from the prediction. In contrast, the correlation decreased slightly with increasing time interval of the prediction an last input observation.

8.3 Recommendation System

The output of our recommendation system is a list that merges both of TripAdvisor and city portal extracted places and averages their scores, in case the place is found in both datasets, and therefore has two scores from both methods.

Observations

As a result, the system will provide a list of touristic hotspots sorted by their scores as shown in the figure 2. The user in this example came from France to Munich with his family during the month of August, his accommodation will be in the district of Maxvorstadt and prefers outdoor places. As it is obvious, the outdoor places are ranked in the first rows with the highest scores.

Trends

Knowing that the current situation of the COVID-19 crisis has set new boundaries between countries, it will be noticeable to see that most of the focus of the recommendation system is going to be on the database of visitors whose origin is from European countries and especially Germany, not forgetting that many places could be closed if the government measures would change again due to another wave of infections. Which might for example enhance the recommendation of outdoor places. While testing the K-Means and score prediction system it was noticeable that the criteria of place type (indoor and outdoor) is an important tool to distinguish the preferences of each user. Another significant citeria for the scoring system is the forecast metrics, which varied through the months after COVID-19. The changes in these predicted values were also decisive which places are most suitable for the visitors.

place_name	place_score
Tierpark Hellabrunn	1.00000
Olympiastadion	0.98160
Olympiapark	0.92774
Olympiaturm	0.73960
English Garden	0.73109
Deutsches Museum	0.71029
Marienplatz	0.59664
Olympiahalle	0.49693
Pinakothek der Moderne	0.33197
Nationaltheater	0.32652
Lenbachhaus	0.29380
Alte Pinakothek	0.27744
Nymphenburg Palace	0.25210
Neue Pinakothek	0.24949
Kleine Olympiahalle	0.23858
Olympia-Eissportzentrum	0.20314
Museum Brandhorst	0.16837
BMW Museum	0.15126
Munich Residenz	0.12747
Staatstheater am Gaertnerplatz	0.12202
Prinzregententheater	0.12065
Muenchner Kammerspiele	0.08657
St-Peter Munich	0.08403
Muenchner Stadtmuseum	0.08180
Bayerisches Nationalmuseum	0.05930
Neues Rathaus Munich	0.05882

Figure 2: Scoring List

9 Discussion

In this chapter the results from the previous section are being discussed and interpreted. For better readability, the three sub-projects are now considered together.

Interpretation of the results

A look at the four correlation tables shows that both models used for the hotspot forecast have almost the same performance. However, since the LSTM network is much easier to optimize, this model should be preferred. The correlation also shows that the prediction is sufficiently accurate. Only the models for two places fell below a value of 0.6.

The same applies to the mean square error. This value is also almost identical for both models. Even within one model, the MSE value varies only slightly from place to place.It can be concluded that the prediction accuracy does not vary much from place to place

This interpretation can be confirmed by the correlation plots. In the plot for the Olympic Tower, which is shown in the Results section, the points form a bisector. In addition, no large outliers are visible. This is also the case with the other correlation plots.

Only at places with few visitors the correlation was more scattered, which indicates a higher model inaccuracy at less frequented places.

Looking at the results of the recommendation system it is obvious that popular places for visitors coming from Munich such as English Garden and Olympiapark are dissimilar

Critical assessment of the results and assumptions

All in all, the model structure is sufficiently accurate for the hotspot forecasting system. However, since the same model structure was used for all public places, it is not possible to optimize it perfectly for all their input.

Proposed Answer to the Research Question

For the first question that concerns the tourism patterns change over the course of the COVID-19 crisis it was possible to create a map which marks the origins of tourists coming from different countries and cities before and after the COVID-19-Crisis with a size of the marker depending on the number of visitors. It is noticeable to see that the amount of visitors coming from Germany has increased after the Corona crisis when the percentage of european visitors has relatively remained the same. As far as for the tourists whose origins are from other countries, the map showed that this specific tourist pattern has decreased due to many reasons such as the new traveling measures, governmental restrictions and health reasons. The touristic guide system could provide detailed information about these patterns and predict the impact of the COVID-19 outbreak on the popularity of each attraction in Munich city depending on various features and reliable sources. One of the obvious changes is the touristic hotspots is the uprise of outdoor places in which people have more degrees of freedom due to the new distance restrictions in closed places. The predicted metrics of popularity was also an essential part for building the personnalized recommendation system that also depends on the user entries and incorporate his preference in the process of ranking the attractioon places. The used datasets and implemented prediction models helped establishing a reliable and robust touristic guide system that delivers up-to-date and accurate results very efficiently.

10 Conclusion

In the course of this document all important points of the applied machine learning project of group 16 were described in detail. First, the database necessary for the implementation of this project was explained. To be able to use the data for training the different machine learning models, the data had to be preprocessed. This part was described in the data preprocessing section. Next, the training of the models and the methods for evaluating the applied artificial neural networks were explained. With the discussion of the obtained results the research question could finally be answered.

Summary of the Results

The in summary results show that since the beginning of the Corona pandemic, most of the tourists in Munich come from the EU and Germany it self. In addition the number of foreign tourists has decreased dramatically. With the Hotspot Prediction System we can predict the popularity and thus the onslaught of tourists for different public places in Munich up to four months. It can be seen that at the beginning of the pandemic, the popularity of all places, indoors and outdoors, fell dramatically. In the course of the pandemic, however, outdoor attractions regained in popularity again.

Future Work

As already mentioned in the discussion, a unique model structure could be designed and optimized for each public place. This should make it possible to slightly increase the prediction accuracy, especially in places with fewer visitors. Another interesting approach would be to predict the relative change compared to the previous month instead of determining the absolute values.

11 Comments to the Group Work Experience

Such large group projects are still a rarity during university studies, especially that in times of Corona the organizational effort had to be increased due to the lack of direct human interaction with the group. Therefore, the situation was a bit unfamiliar at the beginning, but after initial difficulties we adapted to the situation relatively quickly and it turned out to be a great unique experience to work with dedicated and har working people you only knew through a screen. It was also interesting to see how quickly a considerable amount of work could be done due to the advantage we had as a large group. As a team, we have faced some problems, just like any other team working on a big project would, but most importantly we always figured out a solution or a possible way out of it to continue the smooth and continuous work. The contribution of each one of the group did not only mean an individual achievement but also an accomplishment for the whole group, for this purpose were the dedication, productivity, interest and communication of every team member the most valuable elements that made this project succeed.

Parameter	Alte Pinakothek	Bayerisches Nationalmuseum	Bayerisches Staatsoper	Deutsches Museum	Kleine Olympiahalle	Lenbachhaus	Muenchner Kammerspiele	Muenchner Stadtmuseum	Munich Residenz	Museum Brandhorst	Museum Mensch und Natur	Nationaltheater
1. month corr.	0.823	0.565	0.375	0.705	0.497	0.848	0.805	0.624	0.527	0.868	0.895	0.642
2. month corr.	0.779	0.579	0.422	0.626	0.527	0.837	0.808	0.529	0.441	0.89	0.905	0.736
3. month corr.	0.775	0.572	0.506	0.632	0.357	0.786	0.721	0.543	0.517	0.857	0.882	0.732
4. month corr.	0.808	0.599	0.377	0.633	0.532	0.752	0.739	0.569	0.585	0.847	0.898	0.744
MSE	0.438	0.891	0.753	0.736	0.781	0.549	0.748	0.842	0.5	0.409	0.475	0.825

Table 1: one dimensional convolutional neural network evaluation (part I)

Parameter	Neue Pinakothek	Olympia-Eissportzentrum	Olympiahalle	Olympiapark	Olympiastadion	Olympiaturm	Pinakothek der Moderne	Prinzregententheater	Schack galerie	Staatstheater am Gaertnerplatz	Tierpark Hellabrunn
1. month corr.	0.743	0.832	0.676	0.625	0.671	0.779	0.764	0.803	0.64	0.552	0.841
2. month corr.	0.684	0.877	0.599	0.622	0.732	0.791	0.899	0.782	0.597	0.469	0.846
3. month corr.	0.735	0.827	0.509	0.713	0.807	0.852	0.501	0.806	0.58	0.573	0.796
4. month corr.	0.744	0.865	0.495	0.635	0.828	0.804	0.761	0.739	0.605	0.528	0.809
MSE	0.678	0.401	0.793	0.642	0.585	0.483	0.745	0.319	0.567	0.85	0.464

Table 2: one dimensional convolutional neural network evaluation (part II)

Parameter	Alte Pinakothek	Bayerisches Nationalmuseum	Bayerisches Staatsoper	Deutsches Museum	Kleine Olympiahalle	Lenbachhaus	Muenchner Kammerspiele	Muenchner Stadtmuseum	Munich Residenz	Museum Brandhorst	Museum Mensch und Natur	Nationaltheater
1. month corr.	0.843	0.539	0.896	0.706	0.529	0.777	0.709	0.772	0.69	0.621	0.842	0.779
2. month corr.	0.867	0.514	0.922	0.449	0.41	0.752	0.871	0.841	0.739	0.561	0.914	0.789
3. month corr.	0.853	0.503	0.863	0.63	0.522	0.817	0.609	0.859	0.778	0.552	0.845	0.813
4. month corr.	0.861	0.477	0.883	0.731	0.438	0.736	0.837	0.828	0.803	0.509	0.885	0.808
MSE	0.684	0.784	0.643	0.643	0.753	0.621	0.647	0.873	0.815	0.988	0.839	0.892

Table 3: long short term memory neuronal network evaluation (part I)

Parameter	Neue Pinakothek	Olympia-Eissportzentrum	Olympiahalle	Olympiapark	Olympiastadion	Olympiaturm	Pinakothek der Moderne	Prinzregententheater	Schack galerie	Staatstheater am Gaertnerplatz	Tierpark Hellabrunn
1. month corr.	0.436	0.498	0.813	0.865	0.873	0.515	0.562	0.419	0.567	0.457	0.759
2. month corr.	0.542	0.464	0.78	0.863	0.919	0.568	0.62	0.517	0.547	0.518	0.776
3. month corr.	0.493	0.535	0.706	0.91	0.908	0.5	0.621	0.609	0.702	0.899	0.819
4. month corr.	0.766	0.543	0.734	0.884	0.902	0.517	0.576	0.316	0.676	0.898	0.792
MSE	0.894	0.890	0.880	0.845	0.745	0.817	0.788	0.848	0.881	0.881	0.613

Table 4: long short term memory neuronal network evaluation (part II)