RHEIN-WAAL UNIVERSITY OF APPLIED SCIENCES

Machine Learning Based Restaurant Location Analysis in Hamburg

by

Vincent Meyer zu Wickern, Sachin Kumar, Hung Viet Hu

in the

Faculty of Communication and Environment

Geoinformatics, WS 2018/19

February 2019

We declare that this paper and the work presented in it are our own. We confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this paper has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this paper is entirely my own work.
- We have acknowledged all main sources of help.
- Where the paper is based on work done by ourselves jointly with others, we have made clear exactly what was done by others and what we have contributed ourselves.

Signed: _____

Date: _____

Signed: _____

Date: _____

Signed: _____

Date: _____

RHEIN-WAAL UNIVERSITY OF APPLIED SCIENCES

Abstract

Faculty of Communication and Environment

Geoinformatics, WS 2018/19

by Vincent Meyer zu Wickern, Sachin Kumar, Hung Viet Hu

Contents

List of Figures

Abbreviations

ALKIS© Amtliches Ligenschaftskatasterinformationssystem (Authoritative Real Estate Cadastre Information System)

ASCII American Standard Code for Information Interchange

API Application Programming Interface

CPT Central Place Theory

CRS Cordinate Reference system

CSV comma-separated values

ETRS89 European Terrestrial Reference System 1989

JSON JavaScript Object Notation

MetaVer© MetadatenVerbund

OOB out-of-bag

RFR Random Forest Regression

SSE sum of squared errors of the regression model

SST sum of squred errors of the baseline model

SVR Support Vector Regression

WMS Web Map Service

Chapter 1

Introduction

Chapter 2

Elements of a restaurant location analysis

Location plays a critical role in the success or failure of a restaurant business [2] [3] [4]. Not every location may be a suitable location for every kind of restaurant. There are various factors such as demographic values of the restaurant neighborhood, accessibility, visibility and others, which need to be taken into account when selecting the location for opening a new restaurant. To find the best location for a new restaurant, location analysis technique can be used. Some of the elements of restaurant analysis technique are [5] [6]:

1. Demographics

Demographics information of the neighboring population such as age, gender, income, religion, relationship status, environment, and ethnicity have an important effect for potential restaurant owners on choosing a restaurant category like "fast food", "casual dining", "fine dining" or "bar & bistro" [5]. For example, fast food restaurants are most often favored by people with demographic criteria such as an age range between 15 and 35, with a low income, pedestrians and a fast food consumption is usually unplanned and connected to other nearby events like shopping. Fine dining restaurants, in contrast, are favored by people older than 35, couples, high income people, as pre-planned activities and usually mostly by vehicle traffic.

2. Psychographics

Psychographics information such as personality types and personal preferences of consumers support the decision of cuisine types [5]. For example, family-oriented and traditions-oriented consumers usually prefer Italian restaurants, whereas health-conscious consumer regularly choose organic or vegan grill restaurant.

3. Population

Information about the population and the population density in a specific area are necessary to launch any new business or product as defined by the Central Place Theory (CPT) [7]. A location must fulfill the criteria of CPT like range and threshold to be viable location for a new business launch. Range is referred to as the maximum distance that consumers travel for a desired meal and threshold is the minimum population required around the location to start a business or product (restaurant).

4. Customer Activity: Foot & Vehicle Traffic

For a location to be successful, it should have high levels of activities of potential customers around it, like in neighborhood of downtown and tourist places [5]. Foot traffic increases the flow of potential customer walking by a restaurant, whereas vehicle traffic enhances the information to customers who would potentially drive to the restaurant.

5. Competitor Analysis

Analyzing the competitors that are already active in an area can be beneficial in multiple ways [5]. If a location has too many restaurants in the neighborhood, on the one side this means that the regional market is healthy and supportive in that area and the chances of success are high in that area. However, on the other side, it may be difficult to enter the market and attract customers due to the high competition, so sometimes it is better to find a location with less competition, but in which demand still exists.

6. Labor Cost & Minimum Wage

Restaurant environment effects such as the cost of labor, minimum wage and the availability of potential employees vary with different locations [6]. These factors influence the profit and success of businesses to a large extent. Therefore, these pieces of information must be collected and taken into consideration before choosing the location for restaurant.

7. Accessibility and Visibility

For a location to be a potential spot for the opening of a restaurant, it should be easily accessible by foot traffic and car traffic with accessibility to nearby parking area [6]. Moreover, it should not be in a place where the traffic is too high and the restaurant would not be visible by drivers or pedestrians.

8. Proximity to suppliers

Proximity to suppliers is critical, as shipping cost and delivery cost of supplies might be higher than the original net procurement costs of goods, if it not chosen wisely [6]. The optimal choice of a location should minimize these expenses for the business to be successful.

9. Crime Rates

A potential restaurant location should only be finalized after researching in detail about the crime rate and the type of crime in the potential area [6]. Consumers generally do not prefer eating out in areas with high crime rates and it may be a major factor leading to failure of business.

10. Future Growth

Future plans and aspirations should be taken into consideration in the choice of restaurant locations [6]. In the case that the restaurant prospers and gains popularity, it may require an increase in the customer and employee capacity. Moreover, there should be sufficient parking spaces and storage options.

11. Health regulation and zoning

Different parts of cities have different regulations related to health and zones only allowed to certain types of businesses [6]. These regulations should be reviewed and researched properly before finalizing the location.

Chapter 3

Data sources

3.1 Transparency Portal Hamburg

As the first German federal state, Hamburg enacted a transparency law on October 6, 2012 [8]. Opposed to a right to request information, which all citizens had until this date, a new duty to inform the public was laid upon the state's administration offices. All information that would fall under this law, now had to to be published in a freely available standard format on a centered storage of information. The single pieces of information, which would fall under the law, varied highly in precision and the comprehensive term of "geodata" was requested opposed to precise datasets of geodata. A legal interpretation was worked out for all requested points and a plan for the release of geodata was designed consisting of the basic data for measurement admistration and the technical geodata for special administration offices. The transparency law granted a period of two years for the technical implementation.

In October 2014, the "Transparency Portal" (http://transparenz.hamburg.de/) as the major component of the implementation of the transparency law was released [8]. With this portal, the Hamburg citizens have a multitude of data and documents available that was prior only available to Hamburg's administration. One important focus was the release of geodata that was even before the law in preparation for an "Open GeoData" model. In this "Open GeoData" model, geodata was split into two groups of data sets, one group extractable with little effort, but free to the public and expected with a high use, and another group with expected high demand and high revenue on the sale of this data. For this second group of datasets, more effort with new measurements had to be arranged. With the transparency law in place, all datasets were merged into the Transparency Portal and yielded a much higher download count than the count of dataset sales before the portal was active. The Transparency Portal uses a standardized meta datas repository called the MetadatenVerbund (MetaVer©) in collaboration with other German federal states.

3.2 Online restaurant portals

3.2.1 Tripadvisor

tripadvisor is one of the biggest rating portals for travel and travel related businesses, such as restaurants, with cumulated 600 million reviews and opinions until 2017 [9]. This made tripadvisor a potential portal for the analysis of restaurant reviews to extract data from. However, it was found that any scraping, download or copy of the data with automated or manual methods is legally prohibited from tripadvisor, which excluded tripadvisor as a data basis for the analysis of this paper.

3.2.2 Google Places

As Google is the most frequently used search engine on the world and since the research of consumers in the internet before any buying decision icreases, the data processing company Google may have a high influence on purchase decisions [10] [11]. It further provides a service called "Places" for local businesses, among them restaurants, to present themselves with an opportunity of customers to leave reviews, ratings and answers on common questions [12]. With 100 million places, 25 million updates per day and one billion active users per month, the platform may provide insightful data to the restaurant landscape in Hamburg. The Places Application Programming Interface (API) provides an interface for developers to receive data about places, thereof restaurants, that may be used for an analysis.

However, when the API was explored more thoroughly, only the first five reviews of any restaurant could be retrieved and the review count was not extractable. Without the review count, there was information, how well a restaurant has been rated, but not how many customers have, in fact, rated it, which is an essential information. Therefore, Google Places was discarded as a potential source of restaurant information for this paper.

3.2.3 Facebook

Social media is important in the influence to customer choices of products, especially when customers prepare travel or plan the stay at a venue like a hotel or a restaurant [13]. In these cases, online reviews are often taken into consideration. Companies in the hospitality sector therefore have to carefully aim at the benefits of social media tools. Facebook, as one social media platform, has a large customer base with 2.32 billion monthly active users in the fourth quarter of 2018[14]. Out of these reasons, Facebook was considered as one portal to analyze restaurant data from. The Facebook Graph API provides access to content on Facebook for developers [15]. This Graph API is designed for apps that can read and write to Facebook and thereby connects a third-party service with Facebook.

Despite Facebook being a potential source of information of publicly viewable restaurant data, the access to this information was impeded by an app review process. In May 2018, Facebook launched an enhanced developer app review, which made it necessary for developers to verify one's business, sign a supplemental terms contract and provide means of proving the business, such as utility bills or taxi ID numbers [16]. Since this was not available and since there was no intention of creating a business app on Facebook, Facebook as a source of restaurant information was discarde for this paper.

3.2.4 Yelp

Yelp is a popular and widely used social networking site for reviewing and sharing information about local businesses like restaurants, dentists and mechanics [1]. Yelp was founded in 2004 in San Francisco, California, and currently operates in 32 countries. It had an average of 164 million unique visitors every month (via the Yelp app, mobile web and desktop) and had 177 million reviews in 2018. Moreover, it presents information about events, special offers and provides a platform to connect and discuss among yelpers (registered users on yelp). The company Yelp Inc. sells advertisements to local businesses, but they claim that reviews do not get affected by advisements, e.g. in form of added, manipulated, deleted comments. Figure 3.1 shows the statistics for the distribution of user ratings, recommendations and businesses reviewed by category for review across all categories as of 31st December 2018 [1].

From Figure 3.1, it can be seen that the restaurants have maximum reviews among all categories and the rating distribution and recommended distribution shows that yelpers use Yelp to share positive as well as negative experiences. The general, different types of information Yelp collects about restaurant businesses are ratings, reviews, price categories (inexpensive, medium, high, ultra-high), neighborhood, parking facilities, type of meal served (breakfast, brunch, lunch, dinner), smoking areas, reservation possibilities, delivery services and adequacies for children or groups, etc. [1].

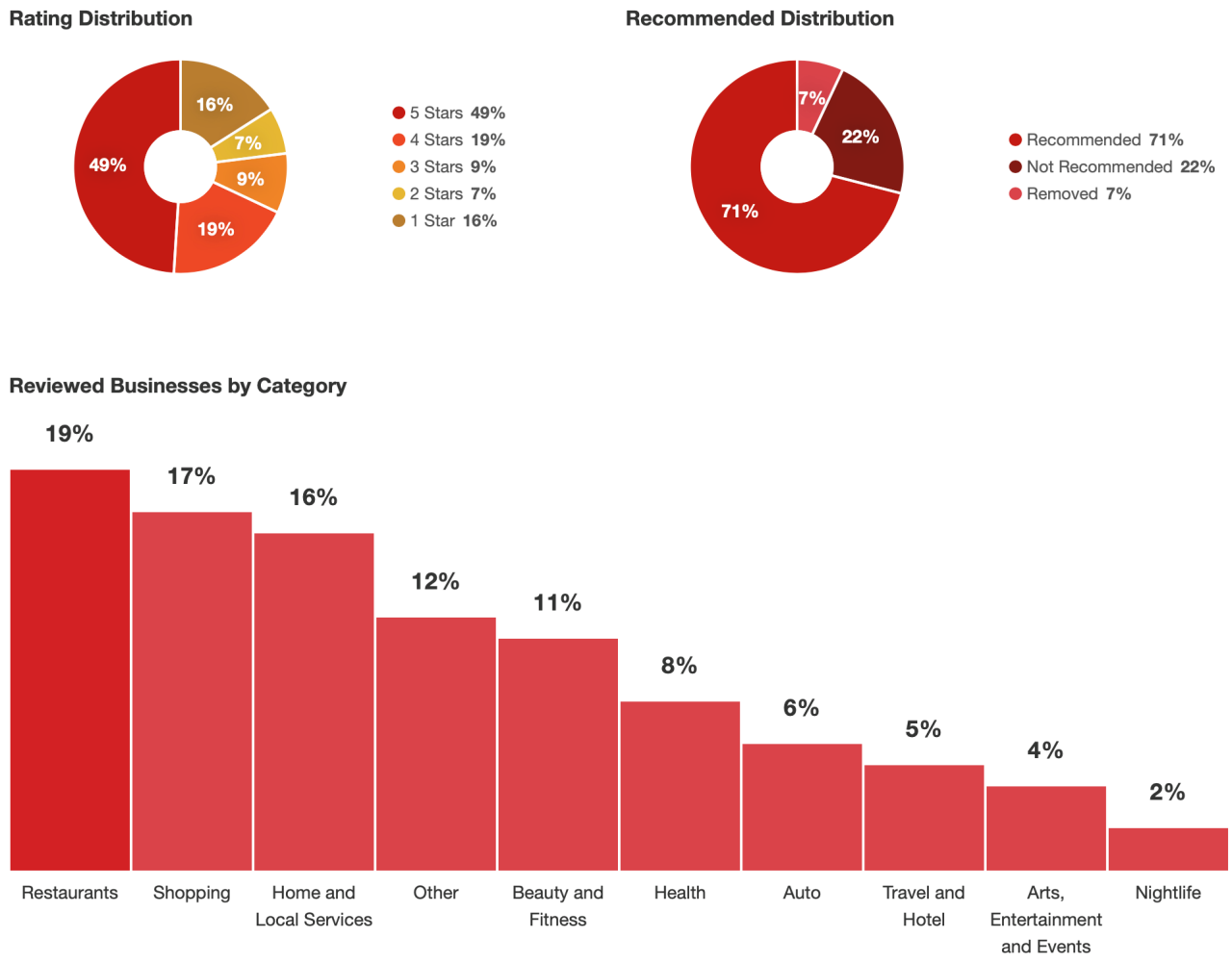

FIGURE 3.1: Statistics on Yelp review according to Yelp Inc. [1]

Chapter 4

Analysis methods

4.1 Random Forest Regression

4.1.1 General description

Random Forest Regression (RFR) is a supervised regression machine learning technique that applies an ensemble learning method [17]. The RFR builds so called "random forests", which are groups of multiple decision trees per random forest. Predicted regression values of RFRs are an average of prediction values from each of the trees that they comprise. The term "random" stems from the fact that individual trees only use a random subset of the observations and the tree splits within a tree as well is based on a random subset of variables. Building these random subsets of observations and attributes, each creating another tree, is conducted in a defined number of iterations. The use of random feature subsets in the RFR technique makes RFR predictions more robust against overfitting than in other techniques, such as support vector machine and neural networks [18]. Moreover, RFR has been found to deliver more accurate predictions in cases, in which the number of variables is higher than the number of observations. In addition, RFR provides two important pieces of information:

1. Variable importance: RFR is able to measure the contribution of variables to predictions [18]. The measure of variable importance can be used to prune less important variables, since the likelihood of overfitting increases with the number of variables.

2. Proximity measure: Proximity measures the closeness of data points to each other and a proximity matrix can be generated [18]. The proximity matrix may then be used to study structure in the data.

4.1.2 Random Forest Regression Algorithm

The steps in the algorithm of RFR can be described as follows [18]:

1. From a given dataset, a defined number of bootstrap samples for trees is drawn.
2. Unpruned regression decision trees for each of the bootstrap samples are developed. Individual trees are based on random subsets of the observations and the splits within the trees are based on random subsets of variables after the defined number of iterations.
3. The dependent variable is predicted for new feature sets by calculating the average of the tree prediction.

4.1.3 Estimation of the random forest error rate

An estimation of the error rate in RFR is obtained with the help of the training data by conducting the following steps [18]:

1. Available data outside of the bootstrap sample (out-of-bag (OOB) data) is predicted at each iteration of bootstrapping.
2. The error rate is calculated by aggregating the OOB predictions. The result of this is known as the OOB estimate of error rate.

4.1.4 Bagging and boosting

Bagging and Boosting are two techniques to construct an ensemble of individual classifiers (trees) [19]. Individual trees of an ensemble are diverse and a higher accuracy of predictions can be obtained by the aggregation and voting of the results of single trees. In the bagging technique, each training set is constructed by replicating the bootstrap sample of the original training set. The single bootstrap samples in this process are called "examples". The example creation in the bagging technique is conducted parallelly. In this process, a defined number of examples are taken from a given training set from the entire population of available data. These examples form a new training set. Multiple of these subsets of the training set are created to build multiple trees in the ensemble. Boosting techniques, in contrast, are based on assigning weights to original training set and adjusting these weights after each regressor prediction. Boosting is a sequential process, in which weights are increased for the examples which are predicted incorrectly and decreased for correct predictions. Moreover, the predictions are less affected by noise in bagging than in boosting. Additionally, bagging constructs diverse trees, only if small variations in the training set cause a large variation in the prediction outputs. RFR uses the

bagging technique in combination with techniques to randomize the internal decisions of the learning algorithm to create diverse trees regression trees.

4.2 Regression performance measures

A common performance measure for the quality of fit of regression models is the "coefficient of determination", which is also called "R-Squared" [20]. This coefficient of determination uses a baseline model, which is a model that consistently predicts the mean of all observations of the dependent variable. This baseline model is a model that a created regression model can be compared to, to see how much more accurate the own predictions were to a poor model.

Following are the calculation of the sum of squared errors of the regression model (SSE), sum of squred errors of the baseline model (SST) and R-squared [20]:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

According to this calculation, R-squared always lies between 0 and 1, if the created regression model yields more accurate predictions than the baseline model [20]. The closer R-squared is to 1, the closer are the predictions by the regression model to the actual values. Additionally, the higher R-squared is, the more variations of the dependent variable are explained by the model.

A weak point of R-squared is that additions of more independent variables never lower the value of R-squared, even for variables with little or no information gain [20]. To cope with this problem, another performance measure called "adjusted R-squared" was introduced, which introduces a negative effect on the measure for the inclusion of ineffective variables. It is calculated as follows [20]:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

n = no of data points

p = no of independent variables in the model

R-squared therefore only improves (moves closer to 1), if significant variables are added to the model and deteriorates (moves closer to 0), if variables are added that are not valuable to the prediction of the dependent variable [20].

Chapter 5

Data extraction

5.1 Hamburg district map

Since usually cities raise important figures in aggregation per admistrative area, in Hamburg being the single city districts, these administrative areas should be imported into QGIS to be able to link single restaurants to an administrative area and hence figures that could be important as dependent variables to predict restaurant success. The borders of the administrative areas are taken from a dataset of the Transparency Portal called "ALKIS Verwaltungsgrenzen Hamburg" [21]. This dataset is available in multiple dataformats, is reported to have a 0% data deficit and a precision of 0.1 meters. It is part of the Amtliches Ligenschaftskatasterinformationssystem (Authoritative Real Estate Cadastre Information System) (ALKIS©), a digital combination of the real estate book information and a real estate map [22].

For this analysis, the GML version of the administrative boundaries were downloaded and imported to QGIS. Loading the administrative boundaries into QGIS, the European Terrestrial Reference System 1989 (ETRS89) Cordinate Reference system (CRS) defined by EPSG:25832 was selected, as was defined in the metadata of the dataset. All following elements that are loaded into QGIS are as well projected in this CRS. The imported boundaries can be seen in figure 5.1.

5.2 Yelp restaurant data extraction

5.2.1 Overview

Yelp provides developers with an API called Yelp Fusion API for accessing the Yelp database. The authentication of API calls is done via a private API key which can be obtained by:

FIGURE 5.1: Administrative boundaries of Hamburg in QGIS

- Registering for a Yelp developer account, and
- Creating an application

An extensive guide on the Yelp API authentication is provided on Yelp's documentation page [23]. For this project, API calls were conducted using Python 3.7.

The extraction of data from the Yelp database consists of:

1. Loading raw data by performing search queries using Yelp's Python API
2. Wrangling data into intermediate formats and storing wrangled data for further analyses

5.2.2 Raw data extraction

5.2.2.1 Retrieving data within a certain area from the Yelp server

The yelpapi Python module [24] must be installed in Python to perform API calls. The obtained private API key must be stored as a Python string:

```
from yelpapi import YelpAPI
api_key = 'pastetheapikeyhere'
yelp_api = YelpAPI(key)
```

Each search query is an https request to the Yelp server. The Python Yelp API exposes this as the following function:

```
yelp_api.search_query()
```

The list of parameters for this function is the same as the https query which is provided by Yelp [25]. For this project, this function was used as:

```
response = yelp_api.search_query(term='restaurants',
 latitude=latitude, longitude=longitude,
limit=50, offset=0, radius = radius)
```

where latitude, longitude and radius defined the centroid of the search area. Note that for each query, only a maximum of 50 restaurants could be returned (default=20). This means that:

- The total number of restaurants had to be extracted from the returned object of the first function call
- The function had to be called multiple times to get data for all the restaurants within the desired search area

This process was built into the get_restaurants(latitude, longitude, radius) function in the Python file yelp_hamburg_load.py.

5.2.2.2 Determining latitudes and longitudes for restaurant searches

To receive data of all restaurants in Hamburg, the city had to be divided into tiles of size 2000 meters by 2000 meters. As Yelp only responded with less or equal than 1,000 restaurants, the tiles were even quartered, when exactly these 1,000 restaurants were sent back. With smaller tiles, all restaurants could be retrieved. The vertices of the tiles were the latitudes and longitudes used by the function get_restaurants() mentioned in section 5.2.2.1. This process consisted of the following steps:

- Pre-determine the maximum latitude, maximum longitude, minimum latitude and minimum longitude of Hamburg. These points were manually picked out from a map of Hamburg in QGIS.
- Create a new empty list of latitudes and a new empty list of longitudes.
- Inner loop: loop from the minimum longitude to the maximum longitude, incrementing by 2000 meters (length of tile's side). Append each new longitude into the list of longitudes.
- Outer loop: loop from the minimum latitude to the maximum latitude, incrementing by 2000 meters (length of tile's side). Append each latitude into the list of latitudes.

Note that the set of pre-determined maximum/minimum latitudes and longitudes were of ETRS89 (EPSG:25832) CRS, while Yelp API calls require WGS84 (EPSG:4326). Therefore, a helper function [transform_coord()] using the Python module pyproj [26] was written to transform the formats of the latitudes and longitudes so that they could properly be passed into the Yelp API calls.

5.2.2.3 Extracting data of all restaurants in Hamburg

As mentioned in section 5.2.2.2, each Yelp search query requires a latitude, longitude and radius to determine the search location. The radius was calculated from the length of the tile's side which is 2000 meters by the following formula:

```
radius = int(math.sqrt(math.pow(TILE_SIZE,2)*2)+1)
```

which was constant during all Yelp API calls.

Each {latitude, longitude} pair to be passed into each Yelp API calls was taken from the lists of latitudes and longitudes created in section 5.2.2.2.

The function get_restaurants(latitude, longitude, radius) was called for each {latitude, longitude} pair until all pair combinations were used.

The result of each call was stored into a JavaScript Object Notation (JSON) file respective to the latitude and longitude used with the help of the Python JSON [27] module. All generated JSON files were subsequently merged into one JSON file for ease of data wrangling.

5.2.3 Data wrangling and storing

As data in the JSON file mentioned in section [link] was un-wrangled, it must be loaded back into Python for processing by:

```
with open('merged_extract.json') as json_data:
  businesses = json.load(json_data)
```

A namedtuple "Restaurant" was created as an object model into which data from the JSON file would be passed into. The data fields for this namedtuple reflects the data retrieved for each restaurant from the search queries. The namedtuple's data fields include:

- epsg25832_latitude

- epsg25832_longitude
- epsg4326_latitude
- epsg4326_longitude
- price rating
- review_count
- is_closed
- zip_code
- category_restaurant_id
- category_alias
- category_title
- category_index

Originally the data loaded from the JSON file was a Python list of (nested) dictionaries. Therefore, data was un-wrangled using the dictionaries' keys. For example, the 'epsg4326_latitude' field for the each instance of the namedtupple Restaurant would be:

```
epsg4326_latitude = biz['coordinates']['latitude']
```

where 'biz' is a dictionary object representing data of one restaurant, which is part of the loaded list of dictionaries from the JSON file.

The list of namedtupples representing restaurants could then be passed into a pandas dataframes and subsequently stored into comma-separated values (CSV) files for further analysis.

5.2.4 Restaurant success calculation

The success of a restaurant is usually measured by financial figures, such as revenue, profit or business growth. However, this information is not publicly available for all small businesses, so a restaurant success has to be assumed from other metrics. Other metrics are important to businesses today are online reviews and ratings. They may influence the success of a restaurant and as well may mirror a restaurant's success. In fact, the rating and reviews a restaurant displays on a portal such as Yelp are their display of success to their customers on the Internet, for which reason the success measures in this paper should combine the two figures of review count and average rating.

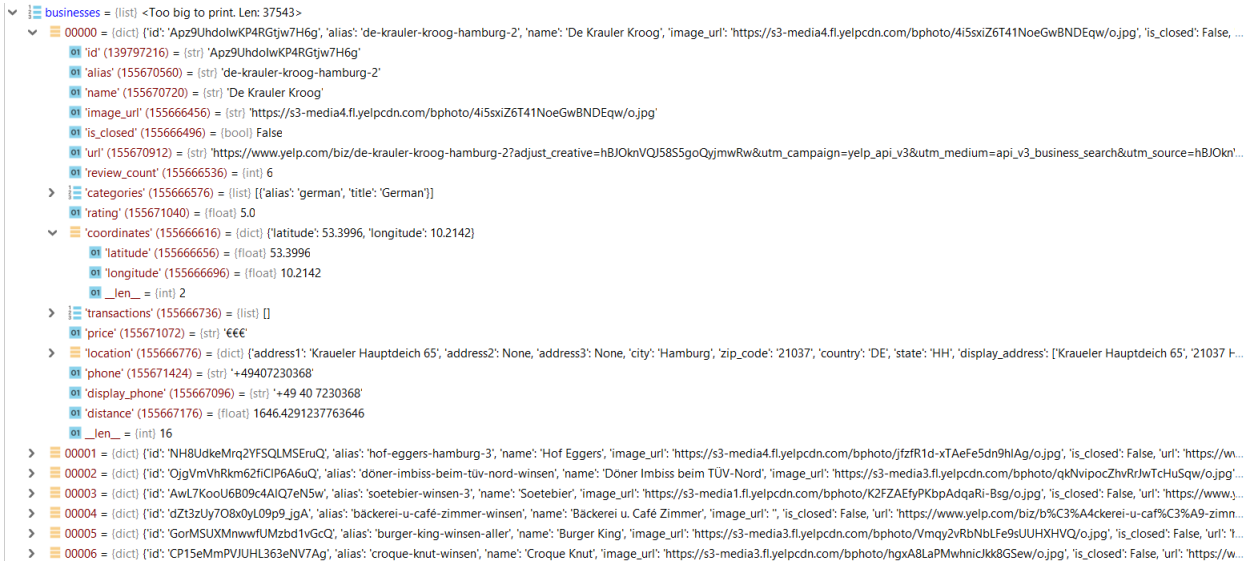

FIGURE 5.2: Un-wrangled data from JSON file

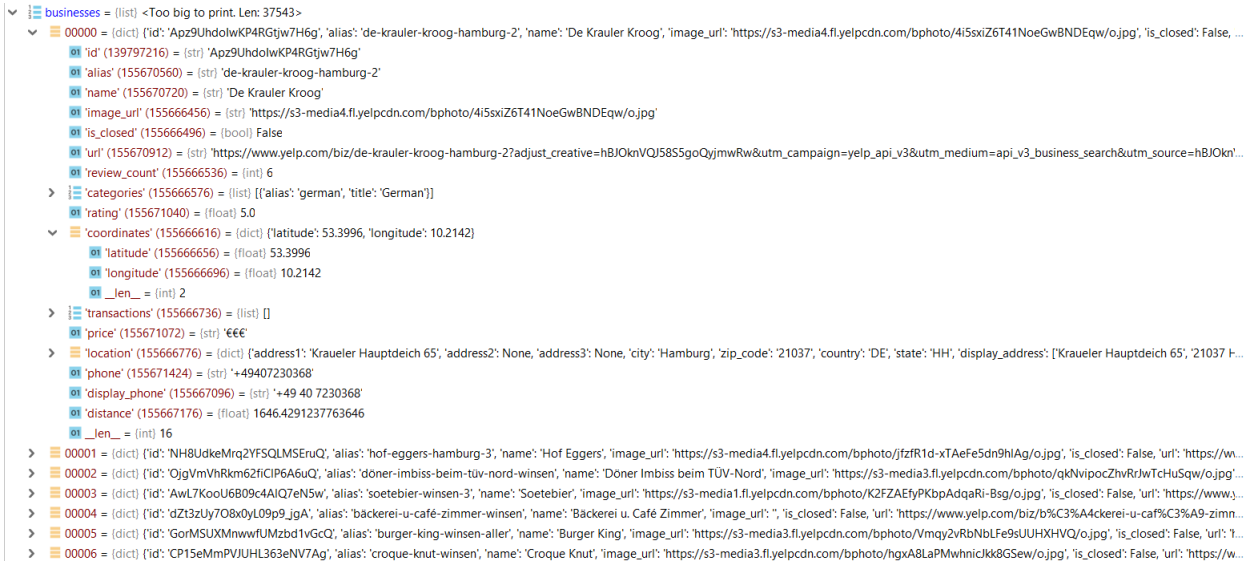

FIGURE 5.3: Wrangled data stored in namedtuple

The average rating and review count have been standardized to set them to a comparable range of values while still keeping the effects of outliers. The standardization has been conducted with the StandardScaler function of the scikitlearn-preprocessing package. Having the average rating and review count in comparable ranges the values were added and saved into a success variable to value them equally in their part of the success. The success, hence, is the sum of the standardized Yelp review count and average rating. In figure 5.4, the distribution of the success values can be seen, they are in the range between ca. 4.2 and 1ca. 13.30.

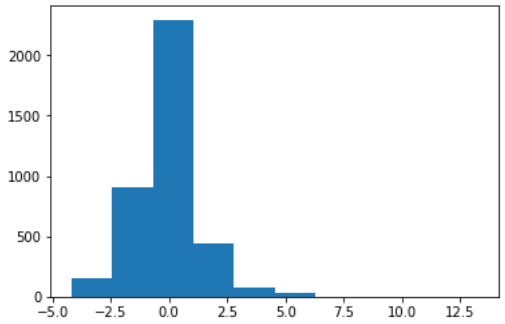

FIGURE 5.4: Distribution of Restaurant Success Values

5.3 City district profiles Hamburg

In this study, it was considered that the city district that a restaurant was located in, may have an influence on the success of this restaurant, e.g. through direct or indirect effects from the demographics in this district. Therefore, "city part profiles" published in the Transparency Portal of Hamburg were taken into account [28]. The dataset presents structure data for all of Hamburg's city parts for the topic areas of population, living, council elections, social structure, infrastructure and traffic. The last dataset was published on the 19.03.2018, but dated back to data from 2016.

The data was available in XLSX-format, in which it was downloaded. Afterwards, the formatting was adjusted to be able to save the content as a CSV. With this CSV, all the necessary information, that would later be used in the analysis, was available as a spreedsheet that could be loaded into Python.

5.4 Proximity to water

As Hamburg is an important port city and almost 10% of the city is covered by the habor [29], water is an important consideration in the analysis of restaurant success. To know the water locations in hamburg is important, because restaurants are regularly not placed on water (restaurant boats are not condidered as potential location candidates for simplicity). Additionally, water may influence the success of a restaurant with customers who may like to sit with a water view. Out of these reasons, the locations of water in Hamburg should be made out and the proximity of each restaurant to the next water location should be calculated.

For the extraction of water location, a geological ground map of Hamburg in the scale 1:5,000 [30] was added as a Web Map Service (WMS) layer into QGIS, as depicted in figure 5.5. The dark blue color on this WMS layer indicates that an area is on water ground, e.g. a river. Since the water spots on WMS layer were not able to be processed as measurable points, yet, these had to be preprocessed before.

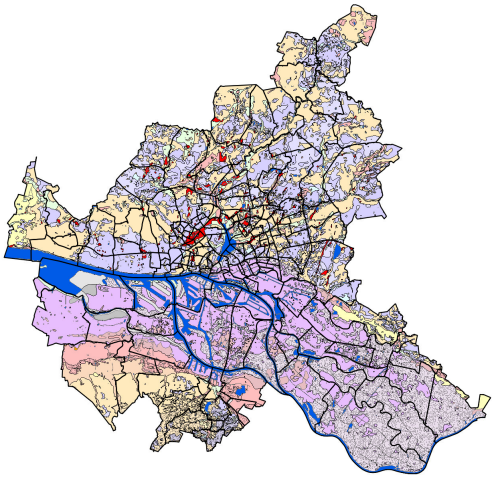

FIGURE 5.5: WMS Layer "Geologische Karte Hamburg"

In the first step, the WMS map was saved as a raster with the "Save as..." option of QGIS. The output was set to a "Rendered Image" in Format "GTiff". The CRS stays ETRS89 and the resolution is first set to 100%, described by entering a 1 into both horizontal and vertical resolution. A new raster file is created.

The newly created raster file had a high resolution and needed to be reduced to be processed further. The resolution was reduced by conducting the QGIS raster conversion function "Translate (Convert format)". In the upcoming dialog, the "outsize" resolution was set to 3% and the output location was set to a new GeoTIFF file. In the created execution code in the buttom of the dialog, the standard "of"-parameter "GMT" had to be changed to "GTIFF". The resolution was after the reolution reduction still adquate to present the water areas.

In the third step, the new and reduced raster file was converted into a grid that would store a certain color value for each of the points on the raster together with their coordinates. This could be conducted by using the same QGIS conversion function "Translate (Convert format)" as in the previous step, however, in the corresponding dialog, the "outsize" resolution was not touched and the ouput format was set to an American Standard Code for Information Interchange (ASCII) gridded XYZ file. In the generated XYZ-grid, a delimited value storage of three attributes per record could be found. The first two attributes present the x and y values on the ETRS89 CRS and the third value a value for the colors with their luminosity is stored. The luminosity "0" is the darkest value in the grid and marks the prior dark blue water locations. This knowledge was used in a python script that was implemented to create a new dataset by looping through every record in the grid and to only keep the records, in which the luminosity would match "0". Thereby only water locations would be part of the new dataset.

The fourth step consisted of the calculation of the distances of each restaurant to the next water spot by using a Python script. In this script, all restaurants and water locations were loaded into the

memory. Then, for each restaurant, the lowest distace to water was calculated by looping through all water points and calculating the Euclidian distance to them. At the end, a minimum water distance was found for each restaurant that was stored together with the restaurant ID in a CSV-file for the use in analysis.

5.5 Restaurant density

As described in chapter 2 on page 2, restaurant density is an important consideration in a selection of a location for a new restaurant. Therefore, this measure was included into the analysis of this paper. A pedestrian usually is willing to walk up to 400 meters to a nearby restaurant [31]. Hence, the restaurant density would present the number of all restaurant, which would be in a range of 400 meters to a location. To include the restaurant density as a feature into the restaurant analysis, a Python script was created, in which the Euclidian distance of all restaurants to all other restaurants were calculated and the ones with a proximity of 400 meters or less were counted to the density of a restaurant.

5.6 Count of criminal cases

The criminal activity near a location of a restaurant can be an important factor in the success of the restaurant, as explained in chapter 2. The state office of criminal investigation in Hamburg uploaded a crime report in 2017, which presents different criminal activities per city district and additionally shows the total count of criminal actvities per city district [32]. This total count of criminal activities was taken from the report and added as a record set for the analysis in the restaurant location analysis.

Chapter 6

Data analysis

6.1 Feature selection and preparation

To prevent overfitting, the number features in the training data should be kept to only the valuable features that support the prediction success in regression. The available data sources and features have been outlined in chapter 5 on page 13.

Apart from the dependent variable represented by the calculated success of a restaurant, included restaurant information from Yelp are the price level of a restaurant represented by one or multiple Euro-signs, and the restaurant category, e.g. a café, Viatnemese restaurant or a Korean grill. To use these two pieces of information, they were discretized and the restaurant category furthermore needed to be binarized. Additionally with the location of a restaurant, the proximity to the next water and restaurant density were calculated as explained in section 5.4 and section 5.5 and added as features for the regressor training.

As explained in section 5.3 and section 5.6, Hamburg provides city profiles with values representing each city district. Considering the factors for a restaurant location analysis explained in chapter 2, the following numbers were chosen to be included into the regressors training for a restaurant depending on the city district that the restaurant was located on:

- Count of criminal cases
- Population
- Share of under 18 year olds
- Share of foreigners
- Population with migration background

- Share of the population with migration background
- Number of households
- Share of one person households
- Share of households with children
- Share of households with single parents
- Population density
- Employment quote
- Share of unemployed people
- Sum of incomes per tax reliable person in EUR
- Prices for properties
- Prices for condominiums
- Share of students in Gymnasium
- Car density

To prevent the effects that numbers in different value ranges would have, the features were standardized.

6.2 Regressor training

In the training of a regressor, a regression technique is fitted to the training data to create a regressor that is able to predict values for unseen sets of features. In this paper, three different types of regression techniques were evaluated for their adequacy on the prediction of the restaurant success - RFR, Support Vector Regression (SVR) and linear regression.

The available data was split into a training set containing 70% of the data records and a test set containing 30% of the training records. Each of the regressors was fitted with the training set and then should predict both the values from the training set as well as the values from the test set. The training set values were fit

6.3 Exploratory data Analysis

6.4 Creation of a recommendation map

Chapter 7

Results and discussion

7.1 Results

7.2 Discussion

Chapter 8

Conclusion

Bibliography

[1] Yelp Inc. An introduction to yelp metrics as of december 31, 2018, 2018. URL https://www.yelp.com/factsheet.

[2] Gwo-Hshiung Tzeng, Mei-Hwa Teng, June-Jye Chen, and Serafim Opricovic. Multicriteria selection for a restaurant location in taipei. *International journal of hospitality management*, 21(2): 171–187, 2002. ISSN 0278-4319.

[3] H. G. Parsa, John T. Self, David Njite, and Tiffany King. Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3):304–322, 2005. ISSN 0010-8804.

[4] Angelo A. Camillo, Daniel J. Connolly, and Woo Gon Kim. Success and failure in northern california: Critical success factors for independent restaurants. *Cornell Hospitality Quarterly*, 49 (4):364–380, 2008. ISSN 1938-9655.

[5] Evan Tarver. How to choose the best restaurant location for your business, 21.04.2017. URL https://fitsmallbusiness.com/choose-a-restaurant-location/.

[6] Webstaurantstore.com. Restaurant location analysis, 25.07.2018. URL https://www.webstaurantstore.com/article/81/restaurant-environmental-analysis.html.

[7] Li-Fei Chen and Chih-Tsung Tsai. Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. *Tourism Management*, 53:197–206, 2016. ISSN 0261-5177.

[8] Roswitha Murjahn and Sascha Tegtmeyer. Open data/transparenzportal hamburg–grundlagen, umsetzung, erfahrungen, auswirkungen. *Zfv–Z Für Geodäsie Geoinformation Landmanagement*, 5(2016):330–335, 2016.

[9] TripAdvisor Media Group. 2017 tripadvisor annual report and notice of 2018 annual meeting and proxy statement, 27.04.2018. URL http://ir.tripadvisor.com/static-files/840c6d1c-9c17-46c9-b52f-3586ead2515f.

[10] Daying Zhao, Bin Fang, Huiying Li, and Qiang Ye. Google search effect on experience product sales and users'motivation to search: Empirical evidence from the hotel industry. *Journal of Electronic Commerce Research*, 19(4):357–369, 2018. ISSN 1938-9027.

[11] Soyeon Shim, Mary Ann Eastlick, Sherry L. Lotz, and Patricia Warrington. An online prepurchase intentions model: the role of intention to search: best overall paper award—the sixth triennial ams/acra retailing conference. *Journal of retailing*, 77(3):397–416, 2001. ISSN 0022-4359.

[12] Places. URL https://cloud.google.com/maps-platform/places/.

[13] Linchi Kwok and Bei Yu. Spreading social media messages on facebook: An analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly*, 54(1):84–94, 2013. ISSN 1938-9655.

[14] statista. Number of monthly active facebook users worldwide as of 4th quarter 2018 (in millions). URL https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

[15] facebook. Graph api. URL https://developers.facebook.com/docs/graph-api.

[16] Konstantinos Papamiltiadis. Enhanced developer app review and graph api 3.0 now live, 2018. URL https://developers.facebook.com/blog/post/2018/05/01/enhanced-developer-app-review-and-graph-api-3.0-now-live/.

[17] Ulrike Grömping. Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009. ISSN 0003-1305. doi: 10.1198/tast.2009.08199.

[18] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3): 18–22, 2002. ISSN 1609-3631.

[19] Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000. ISSN 0885-6125.

[20] R. Devasthali. Coefficient of determination (r-squared) explained, 2018.

[21] Alkis verwaltungsgrenzen hamburg, 28.02.2018. URL http://suche.transparenz.hamburg.de/dataset/alkis-verwaltungsgrenzen-hamburg8?forceWeb=true.

[22] Authoritative real estate cadastre information system (alkis®). URL http://www.adv-online.de/Products/Real-Estate-Cadastre/ALKIS/.

[23] Yelp Inc. Yelp fusion authentication, . URL https://www.yelp.com/developers/documentation/v3/authentication.

[24] Geoffrey Fairchild. yelpapi github, 2018. URL https://github.com/gfairchild/yelpapi.

[25] Yelp Inc. Yelp fusion /businesses/search, . URL https://www.yelp.com/developers/documentation/v3/business_search.

[26] Jeff Whitaker. pyproj 1.9.6, 2018. URL https://pypi.org/project/pyproj/.

[27] Python Software Foundation. json — json encoder and decoder¶. URL https://docs.python.org/3/library/json.html.

[28] Stadtteil-profile hamburg, 2018. URL http://suche.transparenz.hamburg.de/dataset/stadtteil-profile-hamburg2.

[29] Die stadt. URL https://www.hamburgportal.de/die-stadt-hamburg/.

[30] Geologische karte 1:5 000, 11.12.2018. URL http://suche.transparenz.hamburg.de/dataset/geologische-karte-1-5-00011?forceWeb=true.

[31] Yong Yang and Ana V. Diez-Roux. Walking distance by trip purpose and population subgroups. *American Journal of Preventive Medicine*, 43(1):11–19, 2012. ISSN 0749-3797.

[32] Landeskriminalamt Hamburg. Polizeiliche kriminalstatistik 2017: Ausgewählte delikte nach bezirken / stadtteilen, 2017. URL https://www.polizei.hamburg/contentblob/10538308/47b77ab8d33f4c8483d2efd29588ca5f/data/pks2017-stadtteilatlas-do.pdf.