RHEIN-WAAL UNIVERSITY OF APPLIED SCIENCES

Machine Learning Based Restaurant Location Analysis in Hamburg

by

Hung Viet Hu, Sachin Kumar, Vincent Meyer zu Wickern

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 Hung Viet Hu, Sachin Kumar, Vincent Meyer zu Wickern

Contents

List of Figures

Abbreviations

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

ASCII American Standard Code for Information Interchange

CRS Cordinate Reference system

CSV comma-separated values

ETRS89 European Terrestrial Reference System 1989

MetaVer© MetadatenVerbund

SSE sum of squared errors of the regression model

SST sum of squred errors of the baseline model

WMS Web Map Service

Chapter 1

Introduction

Chapter 2

Elements of a restaurant location analysis

Chapter 3

Data Sources

3.1 Transparency Portal Hamburg

As the first German federal state, Hamburg enacted a transparency law on October 6, 2012 [2]. 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 [2]. 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 data repository called the MetadatenVerbund (MetaVer©) in collaboration with other German federal states.

3.2 Online Restaurant Portals

3.2.1 Other potential restaurant portals

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 [3]. 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.

Google

3.2.2 Yelp

3.2.2.1 Yelp Platform

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 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].

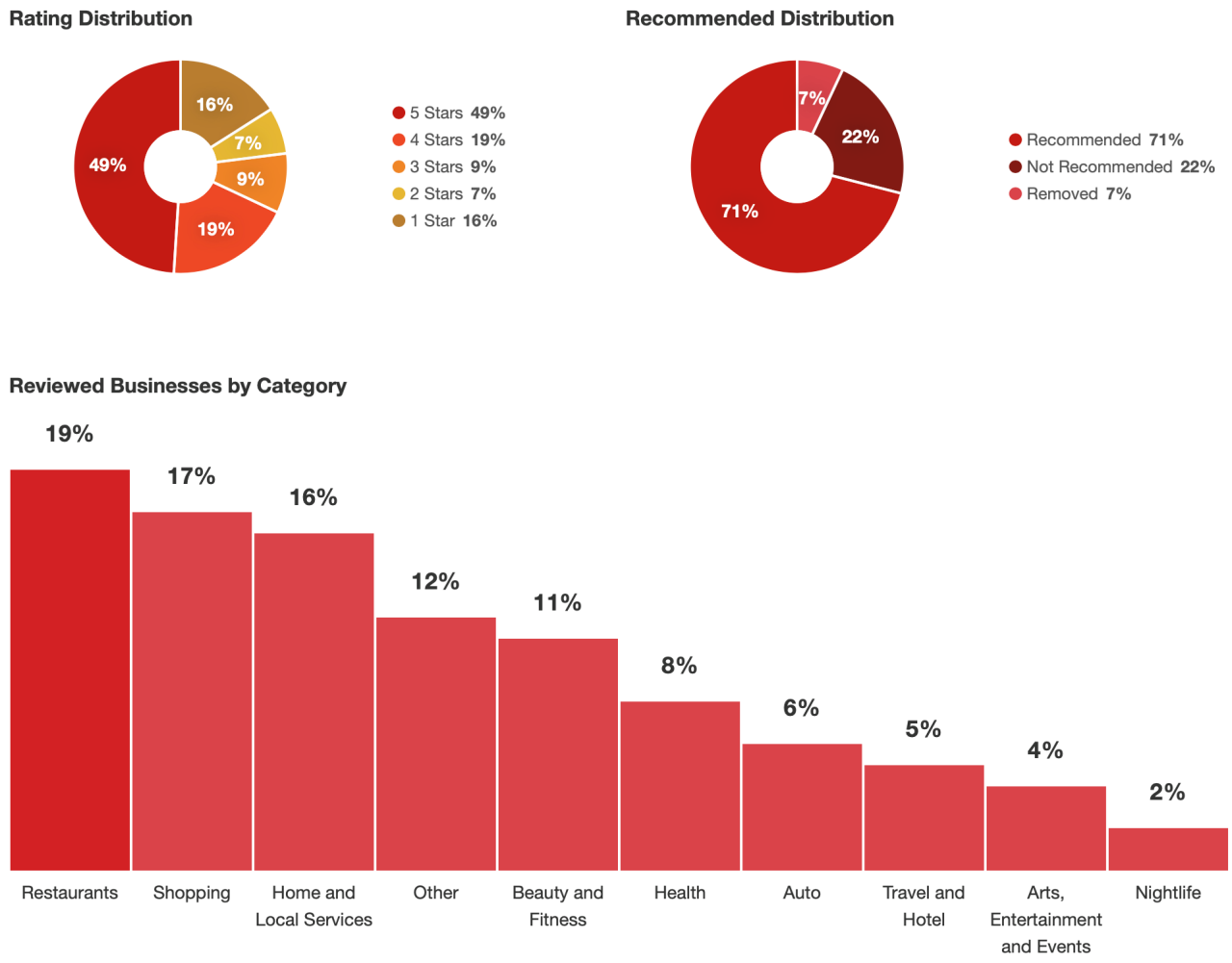

FIGURE 3.1: Statistics on Yelp review according to Yelp Inc. [1]

Chapter 4

Analysis Methods

4.1 Random Forest Regression

4.2 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" [4]. 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 [4]:

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

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

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

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 [4]. 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 [4]. 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 [4]:

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

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 [4].

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" [5]. 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 [6].

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 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

FIGURE 5.1: Administrative boundaries of Hamburg in QGIS

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.

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 5.2, the distribution of the success values can be seen, they are in the range between ca. 4.2 and 1ca. 13.30.

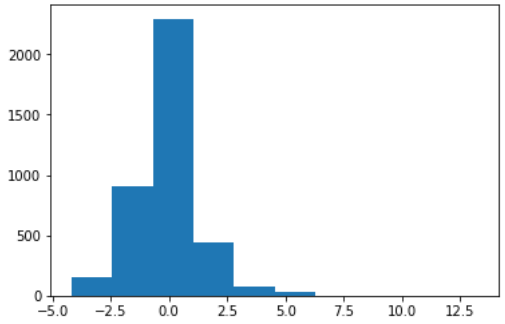

FIGURE 5.2: Distribution of Restaurant Success Values

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 [7]. 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 [8], 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 [9] was added as a Web Map Service (WMS) layer into QGIS, as depicted in figure 5.3. 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.

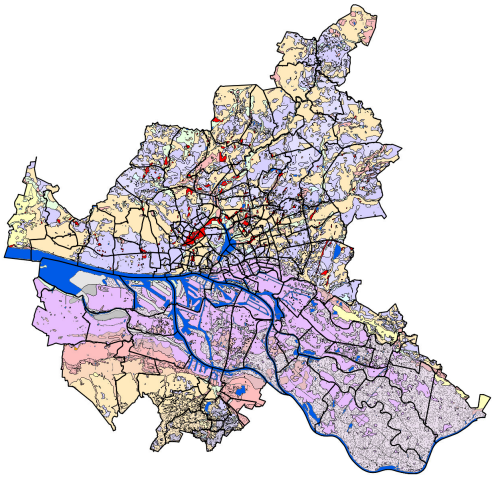

FIGURE 5.3: WMS Layer "Geologische Karte Hamburg"

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 comma-separated values (CSV)-file for the use in analysis.

Chapter 6

Machine Learning

6.1 Exploratory Data Analysis

6.2 Data Preprocessing

6.2.1 Handling of Missing Values

6.2.2 Feature Subset Selection

6.2.3 Dimensionality Reduction

6.3 Data Analysis

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

[3] 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.

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

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

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

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

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

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