Predicting Restaurant Placement and Features

For review rating maximization

By Ilya Smelyanskiy

1. Introduction

1.1. Overview

The restaurant/hospitality industry is highly competitive and diversified in terms of its goods and the manner in which they are presented and sold to restaurant patrons.
It can be narrowly divided into two categories: the commercial restaurant and food service category and the non-commercial food service segment. In the United States for example, the commercial segment accounts for around 80 percent of the market, while the non-commercial segment accounts for the remaining 20 percent.

We are looking at Commercial Restaurant segment of Food Industry, specifically exploring options to open a restaurant in Tallinn, Estonia, that will potentially attract the highest possible customers review and attention.

1.2 Problem

My client needs information on features of highly ranked restaurants in Tallinn to open one himself. He highlights 3 most important questions to him:

1. What is a good place to open a restaurant in the city:
    a. Competition-wise
    b. Visitor-wise
2. What cuisine should a restaurant serve to attract better reviews
3. What properties should restaurant have (internet, wheelchair access, etc.)

2. Data

2.1 Data Sources

For data sources, we will be using an Estonian website, containing information on restaurants in the city Tallinn. The website includes descriptions of cuisine, reviews, important features, etc.

We will obtain information about Tallinn. Thankfully, Estonia is a renowned leader in automation and open data, so anything we might need is located in a register created specifically for devs, like us.

Register is available for public use without limitations, and is provided in XML. We have downloaded addresses by subdistricts in Tallinn from here

As a base for the research, a website www.vabalaud.ee contains some restaurants and data about them.

We will be augmenting this data with Forsquare API, to get additional matches, ratings and misc info, so our information is up to date. We will be matching data by coordinates. Basically, Foursquare will be used to validate data, acquired from www.vabalaud.ee

2.2. Data Preparation

The data parsed from www.vabalaud.ee will be cleaned from special characters, and separated into columns the following way:

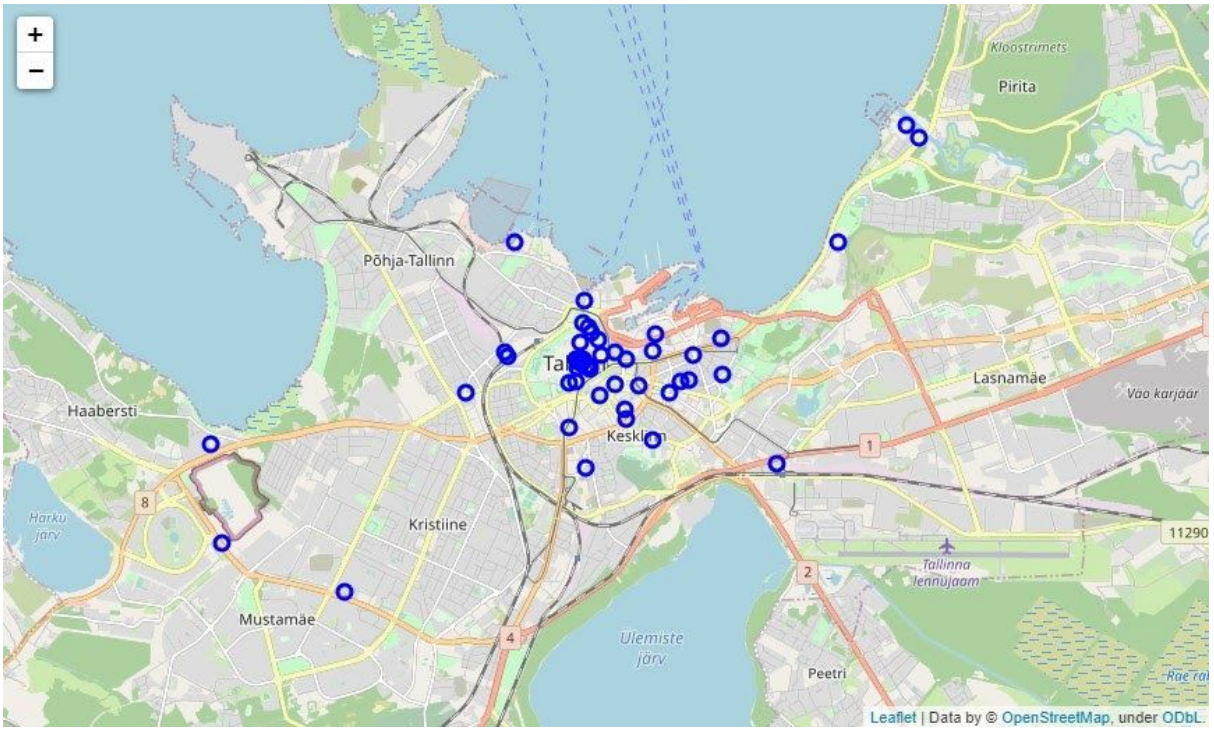
Column name	Description
Restaurant	Restaurant name is displayed here
Details URL	This is url from which data is acquired
Address	Address of the Restaurant
Cuisine	Comma separated types of cuisine
Avg_Bill	Average bill (float), estimated
Additional	List of properties that the venue has (comma separated)
Atmosphere	Atmosphere rating (1-5), float
Food	Food rating (1-5), float
Service	Service rating (1-5), float
Latitude	Latitude of venue
Longitude	Longitude of venue

Data will be extracted via Python's Beautiful Soup library.

3. Exploratory Data Analysis

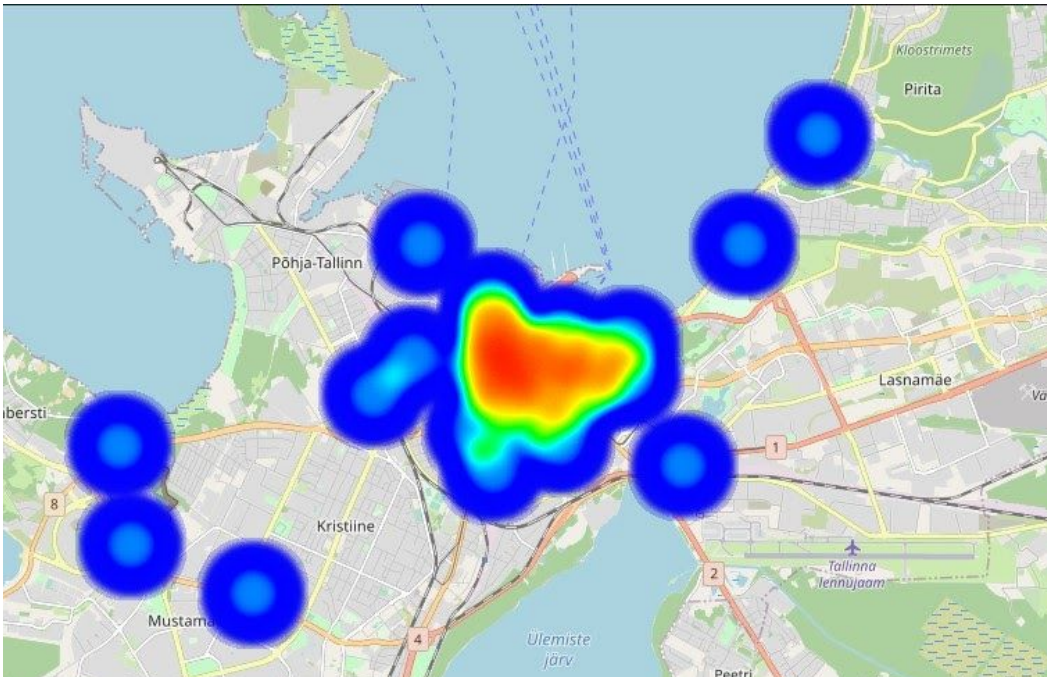
3.1 Restaurants' locations

Once we have acquired the data, we will place it on the map to have a first look.



Img. 1. Restaurants on the map

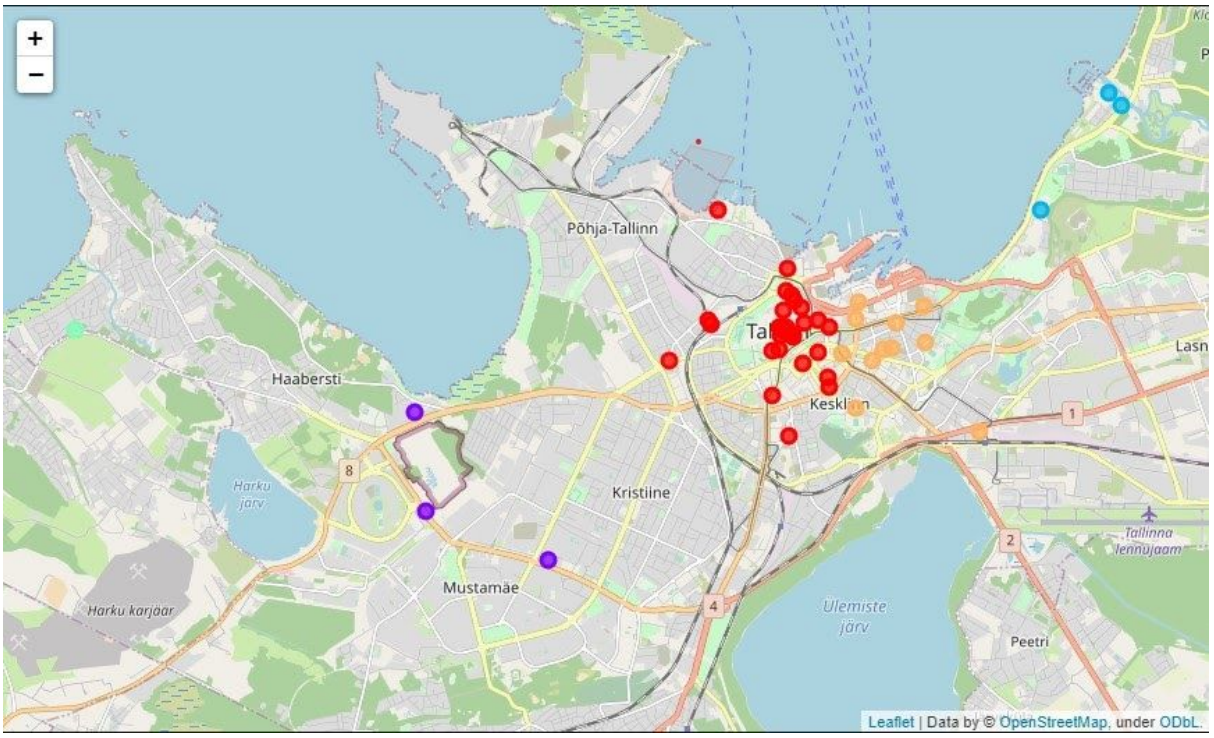
It is noticeable, that the restaurants are located mainly within central town, however, several venues in other parts of the town also exist. We will use heatmap to determine the area of the highest concentration of restaurants:



Unsurprisingly, the entire centre of town is filled with restaurants. It is understandable, since it has the highest amount of tourist attractions.

On the other hand, the entire area of Kristiine lacks a proper restaurant. According to quick Google search, only a few fast-food chains are located in that area.

Additionally, we will run a clustering algorithm, to determine major groups of restaurants.

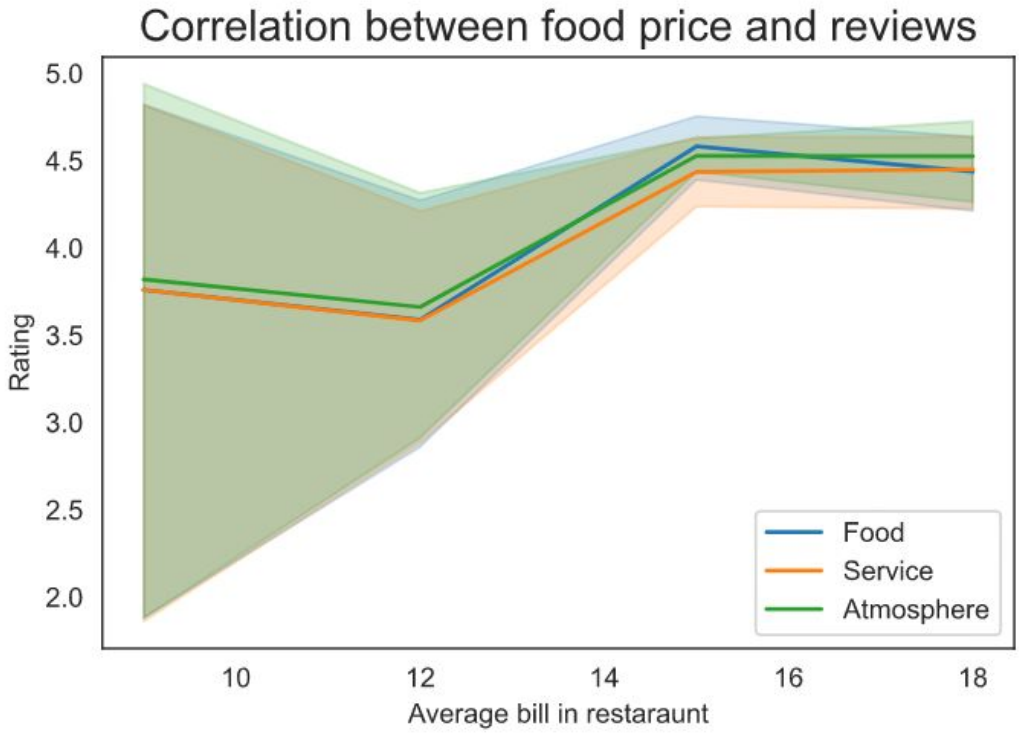


The clustering we see makes a lot of sense. It highlights main groups of venues by coordinates.

1. RED - this is historic town square of Tallinn inside city castle walls. It has a lot of tourists all year round and seems like a good place for a restaurant, but it has a lot of competition.
2. ORANGE - this is a modern business district of the city with several skyscrapers and 5-star hotels located in the area. These restaurants are more fairly spread, while still have a lot of tourists and office workers coming through
3. BLUE - this is Pirreta. It is located at the seaside, near a major yacht club. These restaurants are probably more expensive and flow of visitors is significantly lower here.
4. PURPLE - this is Mustamae, a calm neighbourhood. Fast-food chains and delivery services here are more important than restaurants here.
5. GREEN - This is an outlier, a restaurant located in the less visited side of town. It is relatively new and has little competition.

3.2 How restaurants are graded

After getting the initial data, let's look at how customer satisfaction correlates with good price. Expectedly, more expensive restaurants have higher overall rating. All 3 ratings seem to correlate closely from restaurant to restaurant.



Interesting first observations:

- Expensive restaurants only get high ratings
- Cheaper restaurants aunts have more spread
- Food is rated highest in medium-to-expensive restaurants
- The most expensive restaurants have better atmosphere
- Overall, people don't tend to rate restaurants less than 3.8 stars (or such restaurants close fast)

Since restaurants have a small difference between 3 of their scores, it was decided to use an average score of 3 for further predictions.

4. Modelling

4.1 Goal of Modelling

The goal of machine learning modelling in our case is to find an optimal solution for a restaurant that will receive the best reviews. We have 2 questions to answer in this part:

1. What cuisine is served in high-reviewed restaurants in Tallinn?
2. What kind of additional properties should a restaurant have to get better reviews?

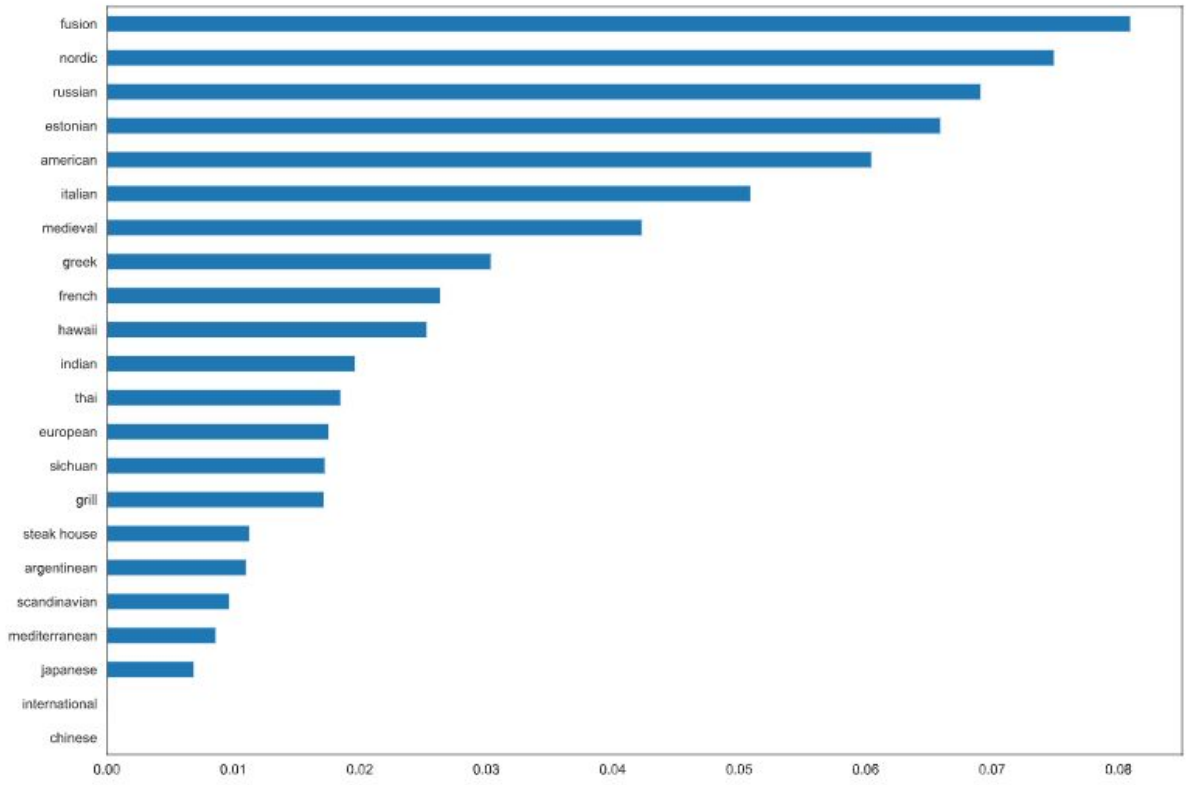
4.2 Model to choose cuisine

To find out what cuisine is most well-received, we will see cuisine's impact as a feature on prediction of positive review.

For this we will teach our model on our tiny dataset of 50 restaurants using Random Forrest classifier, optimizing for average score of the restaurant, based on dummy-values of all kinds of cuisine served in a restaurant (a restaurant can have more than one kind of cuisine)

Overall, quite a few types of cuisine are served in Tallinn's restaurants: Scandinavian, Sichuan, American, Argentinean, Chinese, Estonian, European, French, Fusion, Greek, Grill, Hawaii, Indian, International, Italian, Japanese, Medieval, Mediterranean, Nordic, Steak, Thai.

Considering that we only have 50 restaurants in our dataset, even after augmentation, with Foursquare data, this will not allow for high accuracy.

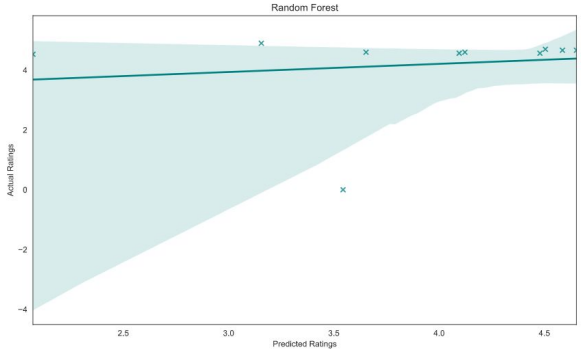


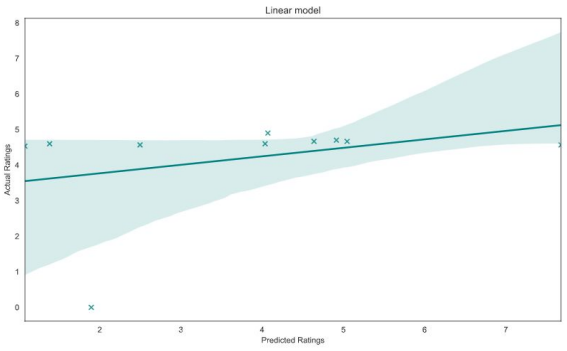
This result was achieved using RF feature importance after training the algorithm. We can clearly see prominence of Nordic, Russian, Estonian, Italian and American foods. This would imply that either the consumers' tastes are conservative based on their culture, or that the dishes in these cuisines are best known to cooks and are produced the best.

4.3 Model to Find Important Features.

We will be using a similar method to the one in the previous step. Main restaurant features will be selected as dummies and fed into a machine learning algorithm. This time

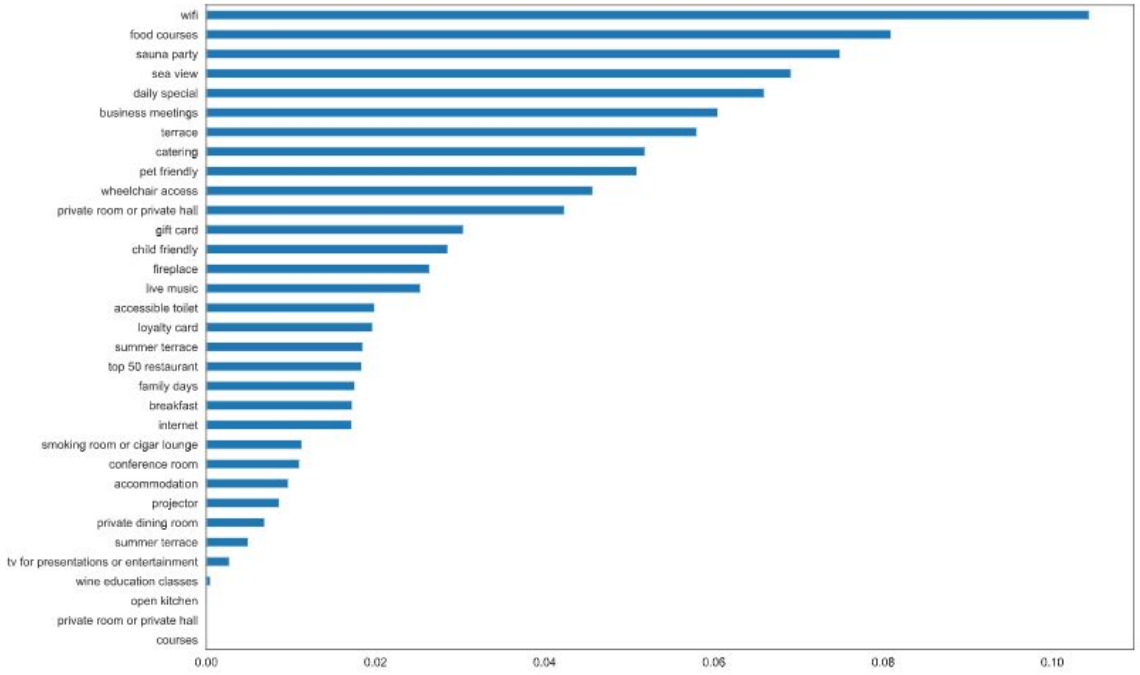
we are trying 3 algorithms: Regression, SVM and Random Forests. Predictions are not very accurate again, due to the limited amount of training data.





However, we are still capable of producing estimation for the most important features. Random Forest algorithm provided best accuracy, so we will use its feature importance function.

The resulting chart of features by importance is as follows:



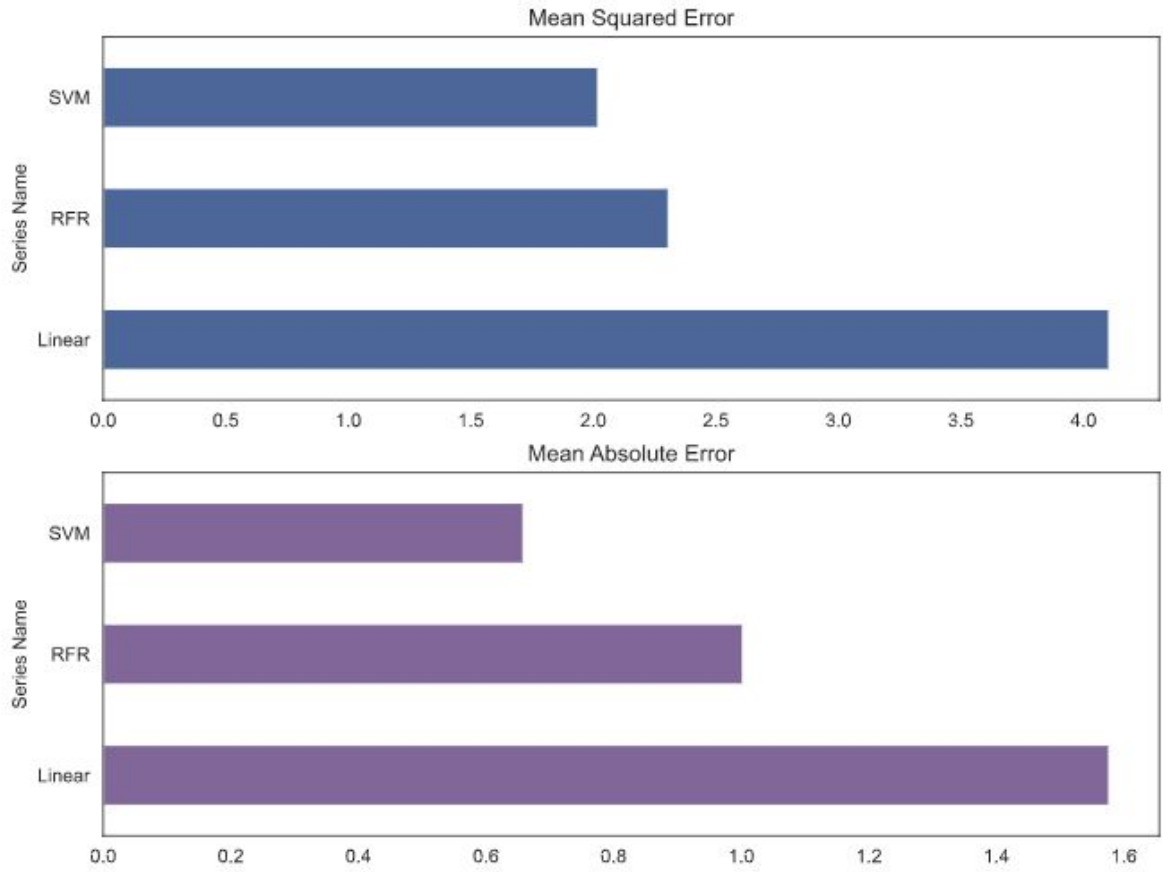
Wi-Fi was a strong predictor of a good review, possibly meaning that restaurants with poor reviews didn't have internet connection.

Another important features were: food courses, sauna parties, sea view, daily specials and business meetings. Terrace, catering and pet-friendly service were also important as predictors of good reviews.

With that in mind, we have the result we were asked to provide.

4.4 Error Measurement

To calculate errors in the ML methods used, we will look at Mean Square Error and Mean Absolute Error.



We can see, that linear regression yielded more error, whereas SVM performed the best.

5. Conclusion

We have provided answers for all the questions raised in the beginning of this research project. We have dived deep into the offerings of the restaurants in the capital of Estonia.

Now we know what food preferences locals and tourists prefer, as well as the kind of places that they would enjoy. We got insight into restaurant placement strategy in Tallinn, and applied several machine learning algorithms to achieve our goals, including K-Means Clustering, SVMs, Random Forrest and Linear Regression.

Unfortunately, our dataset was a little too small, but we managed to get some results nonetheless.

6. Discussion

While we reached our goals, there are a few points I would like to address. Initially I tried to use public transport data from Estonian public data, but they were inconclusive - there was no way to learn the traffic that way. It also wasn't possible with any free APIs.

Additionally, we could've used all restaurants in Estonia or even Baltic to augment our dataset, however, data became very distorted when I attempted to do that.

Finally, Foursquare API unfortunately yielded even fewer results than the website I used to scrape data. This led to mostly scraped data to be used.

There is clearly space for improvement, but at the current stage the project fulfilled its goal both as Capstone for Data Science course, and answered the question of my friend. Let's hope he actually opens his business successfully!