Thesis

### Abstract

### Introduction

Choosing a optimal location for any business is a difficult decision that every entrepreneur faces. As studies show, the location has potentially tremendous effect on revenue. This is particularly important in restaurant industry, where potential customers can be easily tempted by attractive-looking interior. A convenient location is one of the key factors that drive a decision to purchase a meal.

I provide a detailed study of these two factors.

My approach was to create a model explaining restuarants locations using data about population and businesses densities and then assess Variable Importance for each variable.

Modeling is a widely acclaimed method for inference when the relationships are highly complicated and assuming a specific type of data model is impossible.

### Literature overview

#### Business location studies

The location dimension was neglected in mainstream economics for a long time. As Krugman said: *“How did the mainstream cope with spatial issues? By ignoring them”* (Krugman [1997](#ref-krugman1997development)). However, some theories of location were developed through the years. First approaches in the stream of classical economy concentrated on industry and agricuture. The earliest theory concerning location is by von Thunen. (…) His model of agricultural land layed (?) foundations for later works. Theory of industrial location made by Weber (…) concentrated on transportation costs of raw materials and final products. According to the theory, entrepreneurs created their industrial sites in places where the cost of transportation was the lowest.

Later works of Walter Christaller should also be mentioned. He developed a central theory model (…), in which he tried to explain the location of cities and villages across the space. Similar to von Thunen model, a village has one function, that is to create space for exchange of goods produced somewhere else. This assumption was valid in pred- industrial era.

Hotelling’s linear city model (…) is on of classical game theory models. Every firm wants to achieve the best location and attract as many customers as possible. The onovelty of this model is that firms take their competitors’ locations into account. As a result, similar firms are getting very close to each other, and in their interest is to have similar product as the competitors. This phenomenon is visible in retail market, especially bars, restaurants and pharmacies (…).

Hotelling’s theory was then expanded by Salop (…). Main difference was that instead of a strainght line, the street along which businesses competed was a circle with no-end points.

*empirycznie czynniki wpływające na industry i retail*

(Lee and Koutsopoulos [1976](#ref-lee1976locational)) Hypothesised about a relationship between retail stores density and various socio-economic factors in the area. The data was aggregated to a grid and statistical analysis was performed. It was shown that population density does not have a significant influence on stores locations. It was suggested that spatial agglomertion may be a more important factor when making a decision about opening a store.

In an early study, (Rolph [1932](#ref-rolph1932population)) Showed that retail stores location is highly correlated with population density, average income in the area and other factors.

(Esteban-Bravo, Múgica, and Vidal-Sanz [2006](#ref-esteban2006business)) Used questionaires sent directly to randomly selected retail companies in the area to acquire the dat about performance of businesses. They showed nonlinear relationship between business performance and the area it resides in. Firms in high and low businesses density areas performed the worst compared to places in which number of businesses was moderate.

The factors for choosing a site for a facotry and service-based businesses are fundamentally different. For example, in industry transportation costs of raw materials and final products must be taken under serious consideration. Avaliability of a big pool of skilled workforce specialised in a particular industry plays also an important role. (…) analyze an area of the Silicon Valley, USA in context of spatial clustering.

In empirical research, there are two main approaches to estimating business location. One stream of studies concentrates on spatial agglomaration of businesses. It has its roots in Hotelling’s linear city theory. Second stream of studies strives to find factors influencing businesses locations. The reasoning here is that businesses

However it should be stated that these two approaches are highly dependent on each other. The reason of spatial clustering is twofold- one is that (…). However, spatial clustering occurs also because there are good conditions for particular business types in the area.

There is also another groundbound reason why taking account of aglomeration phenomenon is so evident in research studies. Data concerning location is usually very easy to accuire, for instance compared to sales data in different location. No matter how valuable insights one would gather from such information, data of such kind is usually unavaliable to independent researchers. It can be hypothesized that such analyses are made in bigger companies, but the results are proprietary.

(Dubé, Brunelle, and Legros [2016](#ref-dube2016location)) Showed that, in accordance with the classical location theories, businesses in primary sectors tend to be isolated and far from agglomeration center. In contrast, highly advanced manufactuirng and services showed high clustering tendency.

(Van Noort and Reijmer [1999](#ref-van1999location)) Show that small and big businesses should be treated separately in location studies, because motivations and avaliable resources vary considerably. Their study focuses on smaller businesses and is rather qualitative in nature. They claim that SMEs location decision is a short-term and is not a result of a extensive consideration. Also, the smaller the business, the bigger chance is that it will be established in proximity of its owners home.

#### Restaurants location studies

As all of the above authors stated, the results obtained in one city or region should be carefully extrapolated to other areas. Each city has its own specifics, not to mention country’s overall culture and its inhabitants habits.

(Głuchowski, Rasińska, and Czarniecka-Skubina [2017](#ref-gluchowski2017rynek)) Provide an overview of Warsaw restaurants market. They show that the number of restaurants is constantly growing, and Warsaw is a city with the most restaruants in Poland. 3 groups of customers visiting restaurants in Warsaw are shown- people doing this for entertainment purposes (eg. meeting with friends or experiencing new cuisines), tourists visiting Warsaw, and people deciding to take eat outside during the workday, reather than preparing meal at home.

(**???** =1&isAllowed=y) assess the rise and development of *cuilinary tourism*. They claim that a behaviour of traveling to c They provide short overview of restaurants history in modern, post-communism era in Poland, and Warsaw in particular.

From the above studies concerning culinary business it can be stated that until now Poles were not much attracted to traditional cuisine. However, the movement of using traditional, long-forgotten polish recipes and ingredients is on the rise.

One common approach to asserting customers choices in restaurants is using a multi-attribute value theory. In this framework, customers have a set of attributes that they percievie as important for making a decision. These attributes and their personal values are then compared against the attributes of a business. If the assesment is positive, then a purchase is made. This apprach was used in a study by (Auty [1992](#ref-auty1992consumer)), who use segmentation framework to analyze restaurants choices among customers. Various studies have been made and all of them show that importance of various attributes is highly consumer -dependent.

(Johns and Pine [2002](#ref-johns2002consumer)) Provided a review of studies concerning consumer behaviour in restaurant industry.

(Sadahiro [2000](#ref-sadahiro2000pdf)) Analyzed retail location data in Yokohama, Japan using a probability density function estimated.

Restaurants locations can be analyzed from both supply and demand site. Despite the fact that restaurant location decision is made by the owner in the short term, it is consumer’s force and decisions that influence whether particular restaurant will withstand the test of time. Consumer needs and habits are constantly changing and new behavioural patterns can be observed.

(Ayatac and Dokmeci [2017](#ref-ayatac2017location)) Showed how restaurants locations reflected Istanbul’s cultural transformation and opening for Western culture.

(Smith [1983](#ref-smith1983restaurants)) analyzed the location of restaurants in Kitchener-Waterloo.They showed that restaurants locations do not depend on land values in macro scale. However, various strategies are utilized to minimise ifluence of high rent- for example restaruants are located on smaller and less visited streets downtown. Also, restaurants tend to be smaller in high-rent areas compared to similar standard restaurants in other parts of the town. There is an evidence that regular restaurants are mainly located in CBD area, to take advantage of high daytime traffic. The decision of renting a place in a commercial builiding may be leveraged two way- one by avoiding big cost of owning a place, and second by attracting employees from that particular builiding to have lunch there. Smith also emphasizes the importance of zoning regulations the driving factor in restaurants locations in Kitchener-Waterloo.

For example, (Mitchell and Lovingood Jr [1975](#ref-mitchell1975some)) analysed the structure of publicly-funded recreation places based on cetral place theory assumptions.

There is little publically avaliable research on restaurants location. According to Smith, most of the previous research *… has been done under contract for particular restaurant franchises..*, and thus is unavaliable for academic researchers (Smith [1985](#ref-patterns_restaurants)).

Retail location theory is particularly broad. …..

There are 2 main streams of restaruants locations studies. One is mainly concentrated on clustering tendencies in restaurants locations. As proven in numerous studies, (…), various spatial phenomena have been shown to cluster well. This is particularly apparent in locations of some bussinesses categories, as shown by (…).

(Pillsbury [1987](#ref-pillsbury1987hamburger)) Showed that restaurants locations have high clustering tendency. As claimed, *Today, virtually no new restaurant is found outside a cluster of its competitors*. Moreover, restaurants clustering criteria used (socio-economics, ambiance and accessibility) were reflected in restaurants locations. Novelty of this study was not to classify restaurants by their types (fast food, family etc.), but rather the customers’ needs they serve. This is based on the fact that for some types of restaurants (eg. soul food) there is no need for good avaliability, and *journey to dine* becomes an integral part of the dining experience.

(Smith [1985](#ref-patterns_restaurants)) Claim that some restaurants categories (mostly fast food) are highly clustered. Another finding was that population density is exponentialy related to traffic volume in the area. Higher raffic was also correlated with presence of some types of restaurants.

(Binkley and Bales [1998](#ref-binkley1998demand)) Estimated average expenditure for fast food restaurants across American cities using linear model. Among the best predictors were: average fast food price, average grocery price, unemployment rate and number of fast food restaurant in the area. It should also be mentioned that population density was not found to be significant.

(Morland et al. [2002](#ref-neighborhood_food)) Examinated location of food stores and services in south-eastern part of the USA. The main concern of this study was to inspect relation between average income in the neighbourhood and racial structure, and location of food stores and restaurants. They found that in lower-income areas avalioability to high-quality food stores is lower. The same was apparent in mostly black neighbourhoods. Also, the quality of restaurants was bound to average income in the areas.

(Ayatac and Dokmeci [2017](#ref-ayatac2017location)) Examinated spatial distribution of restaurants in Istanbul. In this study, data from 1997 and 2013 were analyzed. Thus, it was possible to analyze temporal dynamics. The influence of GNP per area, population density and distance from sea shore was investigated. First two factors were proven to be significant in both analyzed years. As Istanbul was rapidly developing throughout the years, some changes in spatial structure were observed, eg. restaurants *sprawled* from CBD and historical center to less habitated, suburban areas.

#### Estimation of Complex Spatial Models

Early studies in the area of spatial phenomenas did not account for spatial dependence. The usage of OLS has been common. Later works accounted for spatial lag, however these models were still overly simplistic. In modern studies, more complex models were developed and are in use (…)

Even though studies on spatial regression models are (…), there is a big gap in studies concerning development of spatial classification models. In a canonical book in terms of spatial econometrix, LeSage did not even mention classification task. There are only few publications dedicated to this area. In his study <http://www.ocs.sfu.ca/~ester/papers/KDD-2009-SpatialClassification.final.pdf> (…) developed spatial classification algorithm based on the concept of Voronoi tasselation (…). In his work (<https://pdfs.semanticscholar.org/c9e1/0cf4006690e6f3a3c05a151515d0c5a8ca6d.pdf>) improved decision tree classification algorithm to take into account spatial relations. The main novelty of this study was implementation of this algorithm using GIS software-specific spatial predicates. This was to improve efficiency and velocity of model fitting and predictions. Also, some solutions were proposed to take into account the data of various types (lines, points, polygons). This algorithm is also capable of using information on different levels of aggregation and feed them into decision tree estimation.

Area of advanced non-parametric and machine learning methods is rapidly growing in recent years. Algorithms like Gradient Boosted Models, Random Forest and Support Vacor Machines are state-of-the art solutions when it comes to various prediction tasks. Second concern is usage of spatial dimension in

These algorithms, however, are neglected, when it is important to understand specific process, not only making the best predictions. When it comes to explaining the decisions of algorithms, classic modeling methods like OLS and Logistic Regression are still in large use. Their main advantage, compared to more complex methods, is possibility to quantitavely assess which predictors drive particular decision.

However, because of the fact that complex algorithms cope very well in real-world tasks, efforts are made to create solutions for assesing process of algorithmic decision-making. Another reason for rapid development of Explainable Artificial Intelligence is companies’ need to adjust to European GDPR regulation, specifically right to explanation of algorithm’s decision (<https://www.privacy-regulation.eu/en/r71.htm>). Some of the most important frameworks and algorithms are local interpretable model-agnostic explanations (Ribeiro, Singh, and Guestrin [2016](#ref-lime)), partial dependency plots (Friedman [2001](#ref-friedman2001greedy)) and model-agnostic variable importance assesment(Fisher, Rudin, and Dominici [2018](#ref-fisher2018all)).

Lots of phenomena in real world cannot be explained by a linear or any other simple relationship. Thus, the usage of Machine Learning methods in various areas of research can be expected. Economics is a field in which a vast amount of parametric statistical analysis was introduced and applied.

Most of the practical studies that used spatial classification are standard classification algorithms, fit to spatial data. Some of the studies do not take into account spatial dimension. An example is a geological study of landslide probability made by (Goetz et al. [2015](#ref-goetz2015evaluating)).

Others do, however spatial information is assesed by a primitive method of using geographical coordinates in the model (… lasy?).

(Kanevski et al. [2004](#ref-kanevski2004environmental)) Is one of such studies. It was based on environmental data, however methods developed in the paper can be extrapolated to other areas. A hybrid approach using classical geostatistical tools and 2 machine learning algorithms was used. Main advantage of this method over classic statistics framework is capability of taking into account complex, non-linear relationships. At the same time, the results are still easy to interpret compared to algorithms of which this method consists, that is Artificial Neural Network and Support Vector Machine

(Khan, Ding, and Perrizo [2002](#ref-khan2002k)) Develop an efficient spatial algorithm for classification. Similar to (…link), the main improvement of this study is making a already existing algorithm efficient for spatial datasets. In this approach, a problem of streaming the data and classification *on the fly* is explored.

However, the usage of Random Forest for spatial modeling is not widely populated. Various studies were conducted in natural sciences. (…) analyzed the usage of Random Forest in comparison with Multiple Linear Regression for prediction of carbon mapping in Amazon Forest. They showed that using spatial context with Random Forest improved explained variation by 16%.

Similarly, a (..) study, (Čeh et al. [2018](#ref-vceh2018estimating)) used Ranom Forest and Multiple Regression for apartaments prices prediction. Using the first method, improvement in prediction measured by R^2 was 0.34. In this study, however, spatial dimension was not taken into account.

### Dataset description

In my study I have restricted the analysis to the Warsaw metropoly. The variables included in the dataset are:

* Restaruants locations
* Businesses locations
* Population density
* Bus stops locations
* Roads loactions.

These features come from various sources. Restaurants’ locations were obtained from Zomato website. There were 2341 observations total, but due to incorrect addresses, 72 restaurants were excluded. Population density comes from 2011 GUS National Census (<https://geo.stat.gov.pl/nsp-2011>). The data about businesses was gathered from (…). Location of bus stops and roads was obtained using Open Street Map service.

Although restaurants, businesses and infrastructural features (bus stops and roads) are points data, population density is in a form of an 1km x 1km aggregated grid. Thus, to asses population influence on the presence of restaurants, it was neccesary to convert all variables to the same format. To do this, all variables were binned to a grid in the same resolution as population density data. o The map of restaurants locations (…) shows that there exists high centrality. Also, in regions far from city centre it is visible that restaurants are located in proximity to the largest streets, some of which are exit roads.

Population and business presence are also highly concetrated in the city centre. After binning the points data to a grid it can be seen that both restaurants and businesses locations distributions are highly skewed. Typical power law distribution is observed, with majority of values close to 0 and few observations with extreme values. The population density data is also highly right-skewed, but to a way lower extent than the other two variables.

As shown on the boxplots (…), the subsamples containig and not containing any restaurants are significantly different in terms of business and population locations. Average business count in a grid cell in which the restaurants was present was 116.62, while in regions without restaurants was only 14.53. Similarly, average population density in restaurants’ regions was 5784.74, compared to 1613.98 in regions without restaurants presence. The join-count statistic was performed on restaurants presence data. With p-value< 0.0001, there is evidence that spatial autocorrelation in target variable exists. This means that estimates using non-spatial modeling will be biased, and there is neccessity to take spatial dimension into account.

#### References

### Methods description

#### Ogólny opis zadania

Hipotezy

The goal of this study is to find out what are main factors that drive restaurators’ decisions about whether to open a business in a given location. Specifically, I have tested 2 main hypotheses. One is the influence of business proximity. Importance of this factor would mean that main clients of the restaurants are mainly people during their lunch breaks. The other hypothesis is the influence of population density. This would mean that people go to restaurants in their homes’ proximity in the evenings.

To asses thoroughly the importance of the to factors stated above, I have also included other variables of interest, that can also indirectly influence the restaurants locations. These are proximity of bus stops and total length of roads. These two can be seen as measure of how well is the potential location connected to other parts of the city.

##### Variable importance

(Ishwaran and others [2007](#ref-ishwaran2007variable)) Provide a theoretical assesment of variable importance measures concerning binary regression trees and Random Forest.

(Louppe et al. [2013](#ref-louppe2013understanding)) Is a large extension of work of Ishwaran. The study shows that Mean Decrease in Impurity, a standard alogirthm in Random Forest assesment, is a reliable source of information about the importance of variables. Specifically, the authors prove that the algorithm satisfies basic requirement for variable importance method, that is *“*[*Variable importance*](#variable-importance) *is equal to zero if and only if the variable is irrelevant and it depends only on therelevant variables”*.

(Strobl et al. [2007](#ref-strobl2007bias)) Assess the random forest-based variable importance measurements. They show that some characteristics of independent variables are favoured by an algorithm, and thus a suboptimal subset of features is chosen in the training process. This becomes an issue when there is a mix of categorical and continous features, or when the nominal predictors vary in the number of categories. The authors propose an improved version of random forest algorithm to mitigate that problem.

(Archer and Kimes [2008](#ref-archer2008empirical)) Provide a mean decrease in impurity measure empirical assesment. They show that the measure is reliable assesment of the most important variable in the model. The study is focused on a specific problem, that is gene expression measurement, however the underlying data is of specific type, that is *“…predictor variables are standardized, continuous, and possibly highly correlated ..”*. Thus, one can conclude about the possibility of using that method in other areas of study.

(Grömping [2009](#ref-gromping2009variable)) Show a comparison of variable importance assesment in a regression task by two algorithms, Random Forest and Linear Regression.

One broad class of assesing importance of variables … is through using modeling. This way it is possible to asses influence on target variable in a complex way to mimic true relationships in the data. Also these are non-parametric methods that do not require any assumptions about the underlying process (normal distributions etc.). Fulfilling these requirements are hard in real-world people’ decission processes, as the decision criteria are usually way more complex.

Variable importance in context of modeling is defined as measure to what extent is target variable dependent on considered variable. In the case of measuring influence of business and population location on presence f restaurants, variable importance can be thought of as a measure of that influence.

[…] provide an extensive overview of existing variable importance measures. As they state, *“… variable importance analysis (VIA) techniqueswere developed independently in many disciplines”*.

(Calle and Urrea [2010](#ref-calle2010letter)) Provided a comparison of two Random-Forest-specific variable importance measures- Mean Decrease Accuracy and Mean Decrease Gini. They show that the first measure is highly sensitive to small perturbations in the dataset and generally should be used with caution. They prove that measure based on Gini coefficient is much more robust.

(Janitza, Strobl, and Boulesteix [2013](#ref-janitza2013auc)) Introduce a Area-Under-Curve-based variable importance for using in random forest setttings. They show that the method outperforms the two standard variable importance measures in case of highly imbalanced datasets. The results obtained in balanced classification problems are similar with all 3 methods.

#### Opis Logistic Regression

The research done on the variable importance in logistic regression is not particularly broad. The existing works are only extensions of VI measures in the setting of ordinary least squares. Moreover, there is no dominating method among researchers, as it is in Mean Decrease Impurity in Random Forest setting. However, some measures have been proposed.

(Azen and Traxel [2009](#ref-azen2009using)) Extended a framework of dominance analysis previously developed for linear regression by (Budescu [1993](#ref-budescu1993dominance)).

(Tonidandel and LeBreton [2010](#ref-tonidandel2010determining)) Use a concept of Relative Weights also firstly developed as a OLS tool.

#### MCR

(…) Provide an algorithm for assesing variable importance called *Model Class Reliance*. Its main advantage is that it is a wrapper over existing modeling tools, and thus it is a model-agnostic algorithm. This leads to another important property, that is, one can easily compare the variable importances of two or more algorithms in a meaningful way. Also, a advantage compared to other metrics is that there is no need to fit the model to the data more thatn once. This is especially important in big datasets with many observations and variables, where model training alone, not mention cross-validation procedure, can be a resource consuming task.

The inner wrkings of the algorithms is as follows: first, the model is fit on all variables and a goodness-of-fit measure (AUC for example) is checked. Then, one independent variable is randomly shuffled and the goodness-of-fit is measured using perturbed variable. The process is the repeated for each variable in the dataset. The variable importances can be perceived as the difference between the goodness-of-fit of original dataset and the datasets with perturbed one variable.

In addition, a model in which a dependent (y) variable is shuffled serves as a lower-bound of performance.

The algorithm has been implemented as a part of an R package called DALEX. It is created as a comprehensive set of methods to explain blck-box models using different approaches.

#### Opis 3 metod

#### Spatial weights matrix

As join-count analysis on presence of restaurants shows that there exists poistive spatial autocorrelation, sopatial dimension was taken into account. Neighbourhood was defined with queen criterion, which means that two areas are neighbours if they share at least on edge or vartice. For each variable (including target variable), its spatial equivalent defined as sum of this variable across neighbours was computed. This process is similar to using spatial weights matrix in Geographically Weighted Regression. The reason to include spatial dimension explicitly by adding new variables to the model is due to specification of implementation of modeling algorithms in R software. Aspecially in more advanced models (like Random Forest), adding spatial weights matrix would require rewriting the whole method.

#### Cross validation

Good practice for estimating machine mearning models is using cross-validation. However, this procedure assumes that subsequent folds are independent from each other. For spatial data this posesses a problem, as choosing completely random observations could lead to leakage of information from other folds.

The solution was proposed by (…). He suggests that observations chosen to one fold should be densely located to minimise leakage of information from other folds.

I have used similar approach. I have created (…) folds. To simplify the process of splitting the space, I have used Warsaw’ districts as aggregating units. Each fold consists of 3-4 districts, as shown on map (…)

The choice of districts to include in one fold was arbitrary, the main criterion was to ensure that folds have similar area and the districts inside share borders with each other.

#### Ogólny opis

Some calsses of models have a established method of assesing variable importance.

In my work I have tried 2 methods to asses whether these predictiors are imporant.

I have chosen random forest for two reasons. First, it is a well established model performing well in various prediction tasks (…). Secondly, assesing variable importance is straightforward and model dependent.

I have estimated the model on training data described in section (…). I have used *caret* package as wrapper over randomForest method. In each cross validation phase I have tried different number of predictiors used during each tree fitting. Final accuracy of the model was assesed using AUC criterion, as the target variable is slightly imbalanced (In the training set the share of observations without restaurant is ~67%).

Variable importance in RF models was defined in a introducing publication of this model (…). At each of the splits in training phase, the variable on which to make a split is chosen using Gini Impurity criterion. Importance of given variable is defined as sum of decrease in Gini Impurity in all the splits, in which the variable was used.

Random Forest is a long-established model in classification tasks. During model training, various decision trees are created. Each decision tree is fitted using 1. different subset of observations (obtained from full data set using bootstraping) and 2. different subset of features (variables). This way model overfitting compared to classic decision tree is largely reduced.

In addition to a model with all variables, 3 other Random Forest models were fitted. First one did not include information about business locations. This means that 2 variables were ommited, business count in the area and business count in surrounding areas (containing spatial dimension). Second model was similar, but did not include information about population density. Similarly, 2 variables were ommited. The third model served as a baseline. All variables that are of interest in this study (business count and population density) were ommited. The variables selected were bus stops count and length of roads and their spatial equvalents (these variables summed across all neighbours of target area). To adress spatial autocorrelation fully, spatially lagged target variable was also included (defined as sum of neighbouring areas, in which restaurants were present).

4 models described above were then used to make predictions on previously held out data. Accuracy measured by AUC was reported. Models with all variables and the one without the most interesting variables were treated as upper- and lower-bound for assesing prediction accuracy. Assesing variable importance in this setting is straightforward.

The method used is as follows. Let us consider a classification task, with a set of i (i>2) predictors and x\_i and target variable y. The goal is to asses which variable, x\_1 or x\_2, has bigger influence on the target variable y. We can assume that the best model will be the one containing all variables. Also, the poorest prediction will be given by the model containing all variables except x\_1 and x\_2. Then, models including x\_1 only and x\_2 only will have accuracy somewhere between model 1 and 2.

I have performed similar procedure, but using Logistic Regression as a classifier. The reason is that it is a linear model simple to estimate and use. The predictions are also highly explainable, and thus it is possible to achieve reasoning for every classification easily.

Finally, the last method to asses variable performance is a modified version of my second approach. Similarly, I have estimated 8 models total (4 using Random Forest algorithm and 4 using logistic regression). The difference is in assesing performance of the models. Instead of straightforward AUC comparison on held-out data, I have used resampling method described by (…) and (…caret). First, the models are estimated on training data. Then, instead of using full test set, the observations are sampled with replacement and AUC is computed for each sample set. This way, instead of point estimate of AUC for a given model, one can get multiple values of AUC and obtain estimate of distribution for the results. This method is a improvement compared to simple AUC estimation because it enables to perform statistical inferences on the results. For example, knowing the vector of AUC estimates for two models, one can use t-test to asses equality of estimates between given models, and thus infer about significance of the difference.

### Results

There are 3 variables that were consistently chosen in used methods as the most important: number of businesses in the area, roads length in the area and a variable indicating if there is at least one restaurant in the neighbouring areas. One of the variables chosen for this study for more careful analysis dominated the other, that is population density. This means that for opening and running a restaruant much more important factor is if there are

Spatial autocorrelation also plays an important role in the restaurants location. This is in consistency with the works of (…) who all indicated that restaurants hava a strong clustering tendency.

Archer, Kellie J, and Ryan V Kimes. 2008. “Empirical Characterization of Random Forest Variable Importance Measures.” *Computational Statistics & Data Analysis* 52 (4). Elsevier: 2249–60.

Auty, Susan. 1992. “Consumer Choice and Segmentation in the Restaurant Industry.” *Service Industries Journal* 12 (3). Taylor & Francis: 324–39.

Ayatac, Hatice, and Vedia Dokmeci. 2017. “Location Patterns of Restaurants in Istanbul.” *Current Urban Studies* 5 (02). Scientific Research Publishing: 202.

Azen, Razia, and Nicole Traxel. 2009. “Using Dominance Analysis to Determine Predictor Importance in Logistic Regression.” *Journal of Educational and Behavioral Statistics* 34 (3). Sage Publications Sage CA: Los Angeles, CA: 319–47.

Binkley, James K, and James Bales. 1998. “Demand for Fast Food Across Metropolitan Areas.” *Journal of Restaurant & Foodservice Marketing* 3 (1). Taylor & Francis: 37–50.

Budescu, David V. 1993. “Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression.” *Psychological Bulletin* 114 (3). American Psychological Association: 542.

Calle, M Luz, and Víctor Urrea. 2010. “Letter to the Editor: Stability of Random Forest Importance Measures.” *Briefings in Bioinformatics* 12 (1). Oxford University Press: 86–89.

Čeh, Marjan, Milan Kilibarda, Anka Lisec, and Branislav Bajat. 2018. “Estimating the Performance of Random Forest Versus Multiple Regression for Predicting Prices of the Apartments.” *ISPRS International Journal of Geo-Information* 7 (5). Multidisciplinary Digital Publishing Institute: 168.

Dubé, Jean, Cédric Brunelle, and Diègo Legros. 2016. “Location Theories and Business Location Decision: A Micro-Spatial Investigation in Canada.” *The Review of Regional Studies* 46 (2): 143–70.

Esteban-Bravo, Mercedes, José M Múgica, and Jose M Vidal-Sanz. 2006. “Do Business Density and Variety Determine Retail Performance?”

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2018. “All Models Are Wrong but Many Are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, Using Model Class Reliance.” *arXiv Preprint arXiv:1801.01489*.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*. JSTOR, 1189–1232.

Goetz, JN, Alexander Brenning, H Petschko, and P Leopold. 2015. “Evaluating Machine Learning and Statistical Prediction Techniques for Landslide Susceptibility Modeling.” *Computers & Geosciences* 81. Elsevier: 1–11.

Grömping, Ulrike. 2009. “Variable Importance Assessment in Regression: Linear Regression Versus Random Forest.” *The American Statistician* 63 (4). Taylor & Francis: 308–19.

Głuchowski, Artur, Ewa Rasińska, and Ewa Czarniecka-Skubina. 2017. “Rynek Usług Gastronomicznych W Polsce Na Przykładzie Warszawy.” *Handel Wewnetrzny*, no. 4 (369) Tom II. Instytut Badań Rynku, Konsumpcji i Koniunktur: 118–33.

Ishwaran, Hemant, and others. 2007. “Variable Importance in Binary Regression Trees and Forests.” *Electronic Journal of Statistics* 1. The Institute of Mathematical Statistics and the Bernoulli Society: 519–37.

Janitza, Silke, Carolin Strobl, and Anne-Laure Boulesteix. 2013. “An Auc-Based Permutation Variable Importance Measure for Random Forests.” *BMC Bioinformatics* 14 (1). BioMed Central: 119.

Johns, Nick, and Ray Pine. 2002. “Consumer Behaviour in the Food Service Industry: A Review.” *International Journal of Hospitality Management* 21 (2). Elsevier: 119–34.

Kanevski, M, Roman Parkin, Aleksey Pozdnukhov, Vadim Timonin, Michel Maignan, V Demyanov, and Stéphane Canu. 2004. “Environmental Data Mining and Modeling Based on Machine Learning Algorithms and Geostatistics.” *Environmental Modelling & Software* 19 (9). Elsevier: 845–55.

Khan, Maleq, Qin Ding, and William Perrizo. 2002. “K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees.” In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 517–28. Springer.

Krugman, Paul R. 1997. *Development, Geography, and Economic Theory*. Vol. 6. MIT press.

Lee, Y, and K Koutsopoulos. 1976. “A Locational Analysis of Convenience Food Stores in Metropolitan Denver.” *The Annals of Regional Science* 10 (1). Springer: 104–17.

Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. “Understanding Variable Importances in Forests of Randomized Trees.” In *Advances in Neural Information Processing Systems*, 431–39.

Mitchell, Lisle S, and Paul E Lovingood Jr. 1975. “Some Spatial Aspects of Public Urban Recreation in Columbia, South Carolina.” *Southeastern Geographer* 15 (2). The University of North Carolina Press: 93–101.

Morland, Kimberly, Steve Wing, Ana Diez Roux, and Charles Poole. 2002. “Neighborhood Characteristics Associated with the Location of Food Stores and Food Service Places.” *American Journal of Preventive Medicine* 22 (1). Elsevier: 23–29.

Pillsbury, Richard. 1987. “From Hamburger Alley to Hedgerose Heights: Toward a Model of Restaurant Location Dynamics.” *The Professional Geographer* 39 (3). Taylor & Francis: 326–44.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, ca, Usa, August 13-17, 2016*, 1135–44.

Rolph, Inez K. 1932. “The Population Pattern in Relation to Retail Buying: As Exemplified in Baltimore.” *American Journal of Sociology* 38 (3). University of Chicago Press: 368–76.

Sadahiro, Yukio. 2000. “A Pdf-Based Analysis of the Spatial Structure of Retailing.” *GeoJournal* 52 (3). Springer: 237–52.

Smith, Stephen LJ. 1983. “Restaurants and Dining Out: Geography of a Tourism Business.” *Annals of Tourism Research* 10 (4). Elsevier: 515–49.

———. 1985. “Location Patterns of Urban Restaurants.” *Annals of Tourism Research* 12 (4). Elsevier: 581–602.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8 (1). BioMed Central: 25.

Tonidandel, Scott, and James M LeBreton. 2010. “Determining the Relative Importance of Predictors in Logistic Regression: An Extension of Relative Weight Analysis.” *Organizational Research Methods* 13 (4). Sage Publications Sage CA: Los Angeles, CA: 767–81.

Van Noort, EA, and I Reijmer. 1999. “Location Choice of Smes.” *Blees, J*.