Journal of Contemporary Issues and Thought
ISSN 2232-0032/ e-ISSN 0128-0481/ Vol 13, Issue 2, 2023 (9-22)

Predicting Restaurant Revenue using Machine Learning

Nadiah Hanun Ismail, & Chee-Wooi Hooy*

School of Management, Universiti Sains Malaysia, Penang, Malaysia.
**E-mail: cwhooy@usm.my*

Abstract

This paper studied the restaurant branch's revenue to determine the best strategic location with period of study from 1996 until 2014. On the other hand, the paper examined multiple linear regression, decision tree regression, random forest regression and support vector regression to forecasting approach that will likely generate the highest accuracy during validation process in predicting the revenue. Analysis have resulted that support vector regression gives the lowest of error. Some recommendation been proposed for successful plans toward revenue growth which applicable to adopt in the company.

Keywords: Supervised machine learning; Restaurant revenue prediction; TFI branch

1. Introduction

Over the time, the number of opening the new branch restaurants are keep on increasing for all around the globe. According to National Restaurant Association, restaurant industry in Unites State added a net of 13,817 number of establishments in 2018 which was an increase of 2.2 percent compared to 2017 (National Restaurant Association, 2019). For instance, McDonald's restaurants have 36,525 franchises in 2015 and increase to 38,695 of new franchise on 2019. The growth is around 6% as compared to 2015 and counting till now (McDonald's: Number of Restaurants Worldwide | Statista, 2021).

Increasing of new branch restaurant would give significant benefit to the local. Establishment of new restaurant would therefore contribute to increase the employment availability around the locality. In Unites States itself, currently the restaurant industry employs 15.3 million people at one million locations across the States (Posch et al., 2021). While in Malaysia, the number of employees involve in restaurant sector for 2017 was 958,803 people compared to 891,616 people in 2015 by rate growth of 3.7% per annum and 64.1% of them were employed as fulltime employee (Department of Statistic Malaysia, 2020). In addition, according to Department of Statistic Malaysia, the salaries & wages paid in 2017 was amounted to RM12.2 billion as compared to RM9.7 billion in 2015 by growth rate of 12.5% per year.

Restaurant revenues are not only vital to the owner to itself but also as a whole, the yield would be able to rise the growth of gross domestic product (GDP) of country as well. Total sales in the United States restaurant sector are projected to reach $863 billion in 2019, contributing to 4% of the gross domestic product (Posch et al., 2021). Meanwhile in Malaysia, the Food and Beverage sub-sector recorded RM46.4 billion and contributed 5.4% to the Services Sector GDP (Department of Statistic Malaysia, 2020).

However, to open new branch, the manager needs to decide on what location would give highest revenue in return. Normally, most restaurant manager rely on personal judgement and experience of development teams in selecting the location. Investing into new branch is already consume a lot of money, thus they need to be able to survive in selected new location. Chang and Li (2019) highlighted selecting the strategic location is critical in bringing the customer to the restaurant and yield a good revenue. Failure to select the potential location would be led to the closure of restaurant and consequently would incur huge loss to the restaurant. For instances, Ning Baizura is one of famous singer in Malaysia had open new restaurant but been shut down within 3 months of opening due to not strategic location (Othman, 2019) and same goes to an artist of Zara Zia (Ariffin, 2020).

Nevertheless, by applying mathematical model solution able to give optimal decision on selecting favourable location to the restaurant (Chang and Li, 2019). Hyndman and Athanasopoulos (2018) defined the forecasting as predicting the future development of a particular quantity based on logical methods and current data (Kolkova, 2020). The combination of predicting and machine learning will boost the possible outcome. Align with 20 century trends, many researchers have applied machine learning and algorithm to forecast the sales as address by Bera (2021), Chen et al. (2010), Gogolev and Ozhegov (2019), Posch et al. (2021) and Reynolds et al. (2013).

The main objective of the study is to help the company to make optimal decision on choosing the right new location that yield highest revenue. The aims are to develop and determine an algorithm that effectively predicting the TAB Food Investment's (TFI) branch revenue by comparing 4 different models using the existing restaurant branch data. In general, multiple linear regression, decision tree regression, random forest regression and support vector regression were examined in this study. The chosen algorithms selected based on previous study and a brief explanation of the algorithms would be discussed in methodology. The four forecasting method would then be applied to dataset and conduct the validation test for verification of the outperform method. The smaller the validation test value, the better the predicting quality is.

The chosen outperform model were able to help manager to identify the best strategic location to operate the new restaurant and consequently lead to yield a good revenue which able to cover the cost of building the restaurant significantly. At the same time, the algorithm would facilitate the company to locate the excess of their investment to other important development of business, like training for staff, improve customer satisfaction and focus on innovation. Further, compared with traditional method of discovering the strategic location, the chosen algorithm able to analyse and comparing the multiple new sites. In addition, machine learning takes advantage of the availability of collected restaurant's data while traditional method relies solely on the subjective data. Therefore, human errors can be avoided, and operations can be performed faster than previous methods.

New restaurant outlets incur huge time and capital investments to establish. But failure to choose the right location, the new restaurant may be unable to generate the break even and consequently lead to huge loss in investment and the site closes within a short time. The rest of paper is organized as follows.

We first provide an overview of the previous work in related research areas. The experiments for evaluating proposed models on both train and test dataset would be describe in methodology. Then follow by concise explanation of data exploration and data wrangling in design and analysis section. Next section the result and discussion and highlight the preferred location as well. Finally, the last section would conclude the final project including the limitations and future research directions would be discussed.

2. Literature Review

Several machine learning techniques had been applied to determine the best model in predicting the revenue or sales in various industry sectors. For instances, Hájek and Olej (2010) presents the design of a model for municipal revenue prediction by comparing the support vector machines, feed-forward NN (FFNNs), LRMs for the modelling and use ensemble predictor to improve the prediction performance as well. Support vector machines and FFNNs have proven to be an appropriate method for the prediction of municipal revenue by SVM being the champion (Hájek and Olej, 2010).

For entertainment industry, Ahmad et. al (2020) conduct a study of gaging on Machine Learning Techniques in Movie Revenue Prediction and reveal that multiple linear regression (MLR), support vector machines (SVM), neural networks (NN), random forest (RF), decision tree (DT), logistic regression (LOR), K-nearest neighbor (KNN), and multiple perceptron neural network (MLP) are the most popular algorithms used in movie revenue prediction. The survey point out, the most of the studies showed that variants of neural network algorithm are the most accurate algorithms in movie revenue prediction followed by decision tree and SVM algorithm (Ahmad et al., 2020).

Moving on from short review of non-foodservice industry, now let's see what algorithm previous researcher applied to predict the restaurant revenue. Tanizaki et al. (2019) use Bayesian linear regression, boosted decision tree regression, decision forest regression and stepwise method to forecast the daily number of customers at restaurants and found there was no big difference in the forecasting rate using the method of Bayesian, Decision, and Stepwise, and the forecasting rate of Boosted was a little low (Tanizaki et al., 2019). Reynolds et al. (2013) utilized hierarchical regression model to predict the annual sales volume of the restaurant industry in United State (Reynolds et al., 2013). Hierarchical regression model is a special form of a multiple linear regression analysis in which more variables are added to the model in separate steps called "blocks. While Posch et al. (2021) applied two Bayesian generalized additive in their study for predicting future sales of menu items in restaurants and staff canteens (Posch et al., 2021).

For study of forecasting revenue of a large Russian restaurant chain, Gogolev and Ozhegov (2019) had used linear regression, elastic net (ELNET), support vector regression (SVR) and random forest regression (RF) and discovered SVR and RF regressions outperform results of linear regressions by which the random forest model works better (Gogolev and Ozhegov, 2019). Chen et al. (2010) study used logistic regression model, back-propagation neural networks (BPNNs) model and moving average model to develop method for short term sales forecasting in convenience store for the purposes of controlling the order and managing stock of fresh food. The study reveals BPNN outperforms other methods (Chen et al., 2010).

A study by Hu et al. (2004) exploring the suitable forecasting model to apply for customer counts on-premises buffet restaurant in local casino located at Las Vegas. The researchers have applied 6 different models to tested and evaluated and they were naive model, single moving average, double moving average, exponential smoothing, Holt-Winters and regression model. The result suggested double moving average achieves highest accuracy forecast among other model (Hu et al., 2004). Bera (2021) study a comparison against 4 different model for prediction of restaurant revenue which are base model, weighted model, second stage regression model and ensemble model. The result uncover that second stage regression model produce the highest accuracy which built upon base regression model consists of linear regression, ridge regression, decision tree regressor (Bera, 2021).

3. Methodology

3.1 Forecasting Algorithm

Considering with previous study on algorithm application, several algorithms are employed for revenue prediction purposes. Multiple linear regression, decision tree, random forest and support vector regression (SVR) would be choosing and employed for TFI dataset. The cleaning dataset would be applied with the following machine learning.

1. Multiple linear regression (MLR)
   MLR is a predictive algorithm used in predicting continuous data based on previous training data. Basically, MLR involves estimating the relationship between dependent variables and more than one independent variable. In this case, it is between restaurant revenue of TFI data.
2. Support vector regression (SVR)
   Support vector machine is a well-known supervised learning algorithm used for solving classification and regression problems and called support vector regression when used for regression. SVM are work by finding a hyperplane among 'n' features that best divides the data points. Hence it not effected by outlier and used to minimize an error function.
3. Random forest regression (RF)
   Random forest is a decision tree-based ensemble learning algorithm for classification and regression tasks. It is an ensemble algorithm because it uses more than one decision tree within itself. Random Forest algorithm works by creating multiple decision trees and merges them together to get more accurate and stable prediction. Through this, it solves the overfitting problem (Ahmad et al., 2020).
4. Decision tree regression
   Decision tree is a graphical presenting of all possible solution to a decision based on certain conditions. When the target variable is continuous then called as regression tree. It required less data cleaning and less significant on cases of missing values and outliers.

3.2 Evaluation Metric

There are plenty of forecast accuracy measures and the most widely used methods by Hyndman & Athanasopoulos (2018) at present are root mean squared error (RMSE) and Mean absolute error (MAE) (Kolkova, 2020). Pereira and Cerqueira (2021) and Hu et al. (2014) applied RMSE while Kolkova (2020) and Pereira (2016) have applied both MAE and RMSE to evaluate their model. RMSE is a measure of how close a fitted line is to data points. It is the standard deviation of the prediction errors (residuals). Mean absolute error (MAE) is a measure of the average magnitude of prediction errors ignoring their direction. Hence, the evaluation metrics used in this study would be MAE and RMSE to simplify the comparison of accuracy error among different algorithms.

4. Design and Analysis

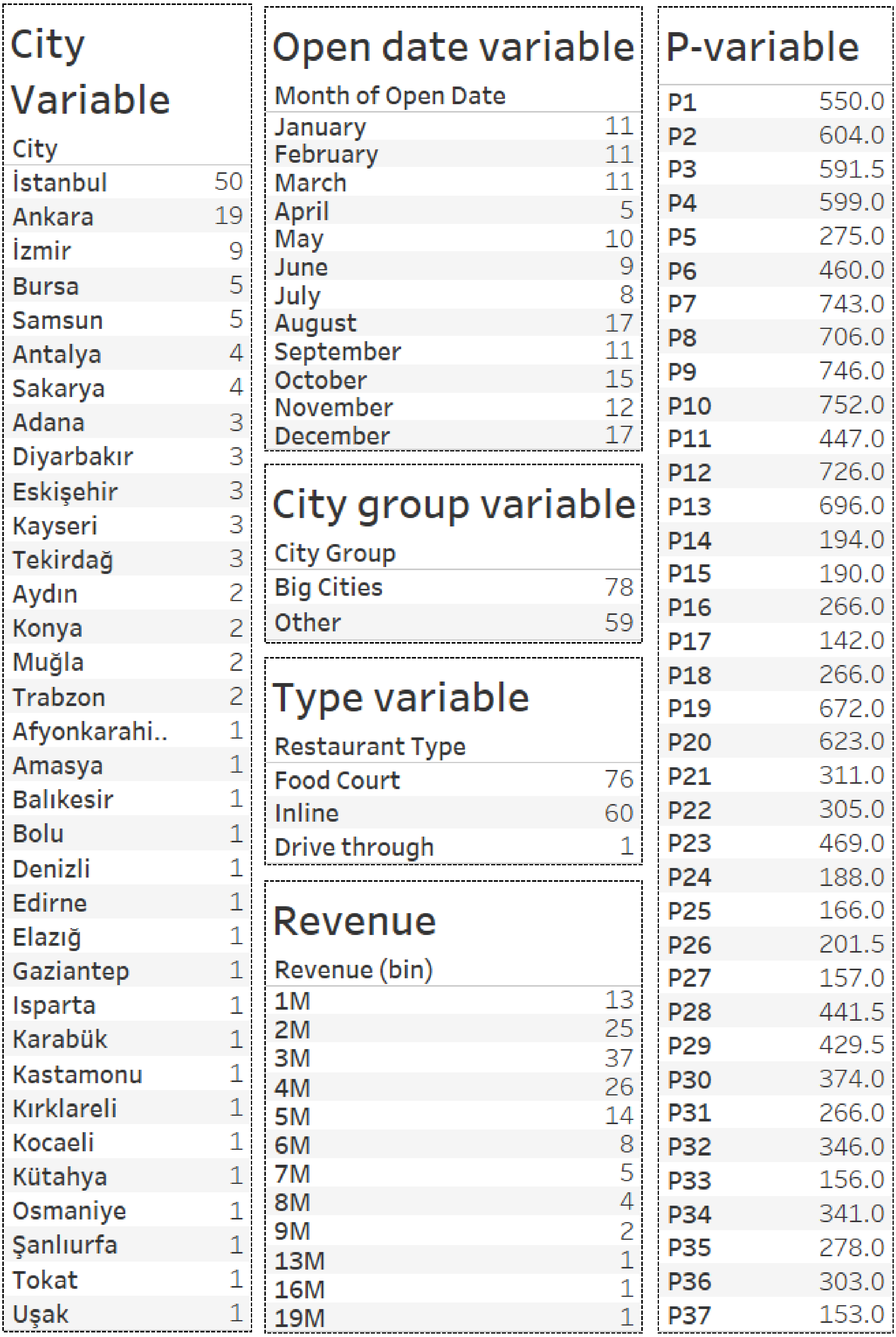
4.1 Data Exploration

The dataset was acquired from Kaggle (Restaurant Revenue Prediction | Kaggle, 2015) and it was based on TAB Food Investment (TFI)'s data located in Turkey. The data were split by training and test set. Training set consists of 137 samples of restaurant and 43 variables while test set contains of 100,000 samples of restaurants and 42 variables. Compared to test set, training set provided with actual revenue amount which classify as target variable in this study. It is interesting to see that the testing set have far more samples than training dataset. The following is a list of the 43 variables available:

1. ID: Restaurant ID
2. Open Date: Date that the restaurant opened in the format M/D/Y
3. City: The city name that the restaurant resides in
4. City Group: The type of city can be either big cities or other
5. Type: The type of the restaurant where FC - Food Court, IL - Inline, DT - Drive through and MB - Mobile
6. P-Variables (P1 until P37): Obfuscated variables consists of three categories: Demographic data, e.g., population, age, gender; Real estate data e.g., car park availability and front facade; Commercial data e.g., points of interest including schools, banks, and other vendors, etc. It is unknown if each variable contains a combination of the three categories or are mutually exclusive.
7. Revenue: Annual (transformed) revenue of a restaurant in a given year and is the target to be predicted

The majority of data are coming from obfuscated variables by which the provider did not give more details about each of the P-Variables.

The exploration of variables from both training and test dataset were conducted by using simple Tableau visualization. Tableau presents appropriate graphical presentation and information effectively by applying the automatic presentation functionality that is intuitive and predictable (Mackinlay et al., 2007). Based on TFI's train dataset, I have designed two different methods of visualization, the first one is descriptive table (Figure 1) and second is graphical visualization (Figure 2) and both performed using Tableau. The purpose is to proof the statement by Mackinlay et al. (2007). Comparing both visualization methods, Figure 2 (graphical visualization) is easier to read and interpret the information given with only a glance. In addition, Figure 2 presenting with appeal visualization and able to attract the reader to captured and understands what the visualization tries to tell. From this onward, all the data would be presented using graphical visualization and next is the explanation of the data.



City Variable

City	
İstanbul	50
Ankara	19
İzmir	9
Bursa	5
Samsun	5
Antalya	4
Sakarya	4
Adana	3
Diyarbakır	3
Eskişehir	3
Kayseri	3
Tekirdağ	3
Aydın	2
Konya	2
Muğla	2
Trabzon	2
Afyonkarahi..	1
Amasya	1
Balıkesir	1
Bolu	1
Denizli	1
Edirne	1
Elazığ	1
Gaziantep	1
Isparta	1
Karabük	1
Kastamonu	1
Kırklareli	1
Kocaeli	1
Kütahya	1
Osmaniye	1
Şanlıurfa	1
Tokat	1
Uşak	1

Open date variable

Month of Open Date	
January	11
February	11
March	11
April	5
May	10
June	9
July	8
August	17
September	11
October	15
November	12
December	17

City group variable

City Group	
Big Cities	78
Other	59

Type variable

Restaurant Type	
Food Court	76
Inline	60
Drive through	1

Revenue

Revenue (bin)	
1M	13
2M	25
3M	37
4M	26
5M	14
6M	8
7M	5
8M	4
9M	2
13M	1
16M	1
19M	1

P-variable

P1	550.0
P2	604.0
P3	591.5
P4	599.0
P5	275.0
P6	460.0
P7	743.0
P8	706.0
P9	746.0
P10	752.0
P11	447.0
P12	726.0
P13	696.0
P14	194.0
P15	190.0
P16	266.0
P17	142.0
P18	266.0
P19	672.0
P20	623.0
P21	311.0
P22	305.0
P23	469.0
P24	188.0
P25	166.0
P26	201.5
P27	157.0
P28	441.5
P29	429.5
P30	374.0
P31	266.0
P32	346.0
P33	156.0
P34	341.0
P35	278.0
P36	303.0
P37	153.0

Figure 1: Descriptive table of train's variables

The graphical visualization of variables for both train and test dataset were presenting in Figure 2 and Figure 3 respectively. From the observation, majority of new restaurant for both training and test dataset were open in August and December. For city variable, the top three preferred locations are Istanbul then followed by Ankara and Izmir in both datasets. In term of city group, big cities have the highest frequency for train dataset compared to test data, it was others category. The type of restaurant indicates food court and inline are the largest occurrence while drive through and mobile shown less in both datasets. Based on analysis of restaurant type, there is no mobile category presenting in training dataset and possibly making it challenging to predict when no weights are existing. However small portion were count in test data and prone to cause minor percentage of error.

The bar chart of P- variable presenting that P10 and P9 were the main contribution of train data while P10 and P7 were the biggest value for the test data. Revenue was plotted using histogram to present the distribution of data. Apparently, revenue variable was long skewed to the right due presenting of outlier. The outlier occurred might be due to several restaurants reporting high revenue or simply an error of set value. Revenue was not available in the test dataset.

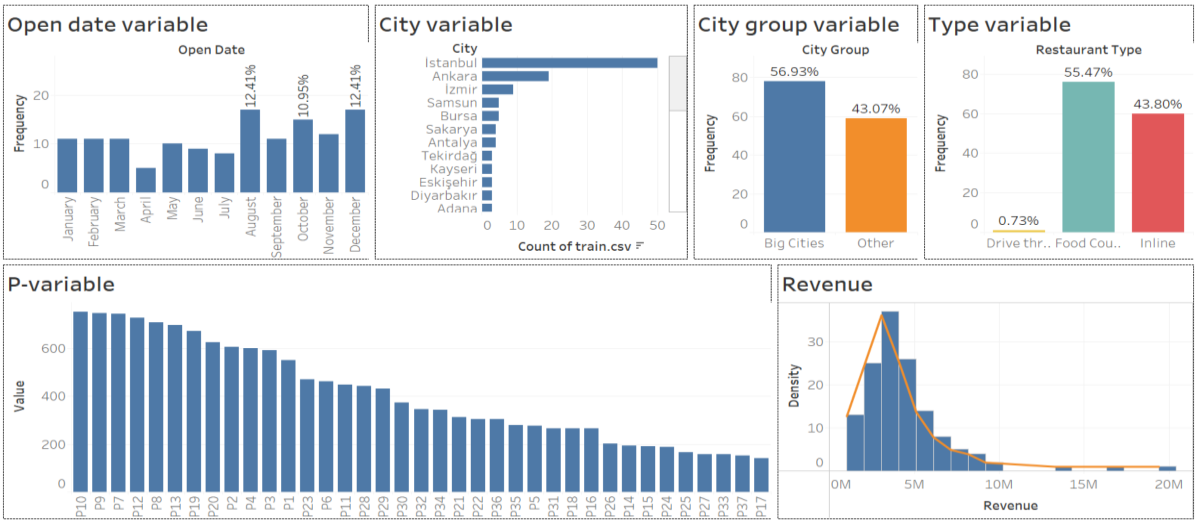
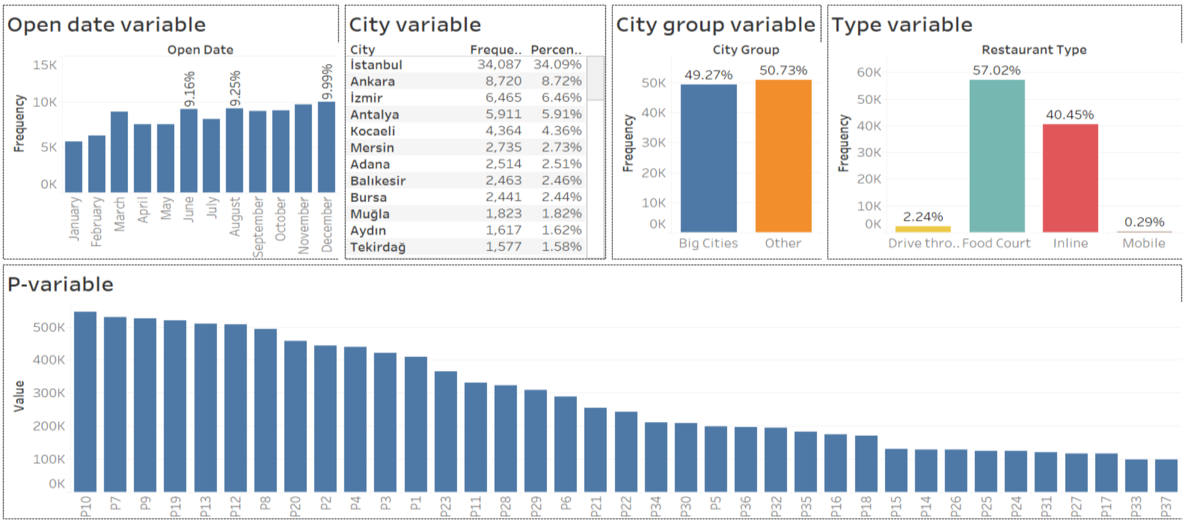


Figure 2: Variables that contains in train dataset



City	Freque..	Percen..
İstanbul	34,087	34.09%
Ankara	8,720	8.72%
İzmir	6,465	6.46%
Antalya	5,911	5.91%
Kocaeli	4,364	4.36%
Mersin	2,735	2.73%
Adana	2,514	2.51%
Balıkesir	2,463	2.46%
Bursa	2,441	2.44%
Muğla	1,823	1.82%
Aydın	1,617	1.62%
Tekirdağ	1,577	1.58%

Figure 3: Variables that contains in test dataset

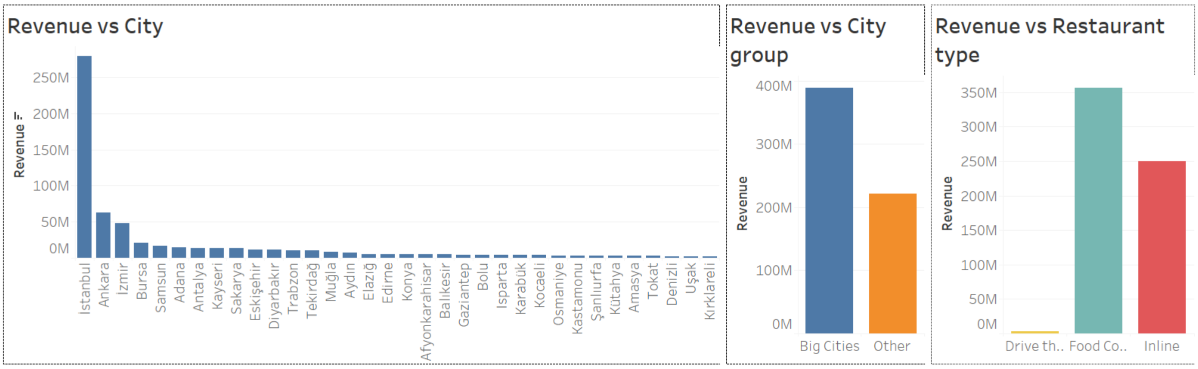


Figure 4: Visualization of relationship between revenue with city, city group and restaurant type for train dataset

Relationship between revenue and city indicate Istanbul surplus the average revenue of the other cities, then followed by Ankara and Izmir. As expected, these three cities were the most preferred cities to open new restaurant as contain in train dataset. For correlation between revenue and city, operating new restaurant in big cities would have high possibility to yield good revenue. In similar way, restaurant that choose to design as food court type tend to generate high revenue compared to inline and drive through.

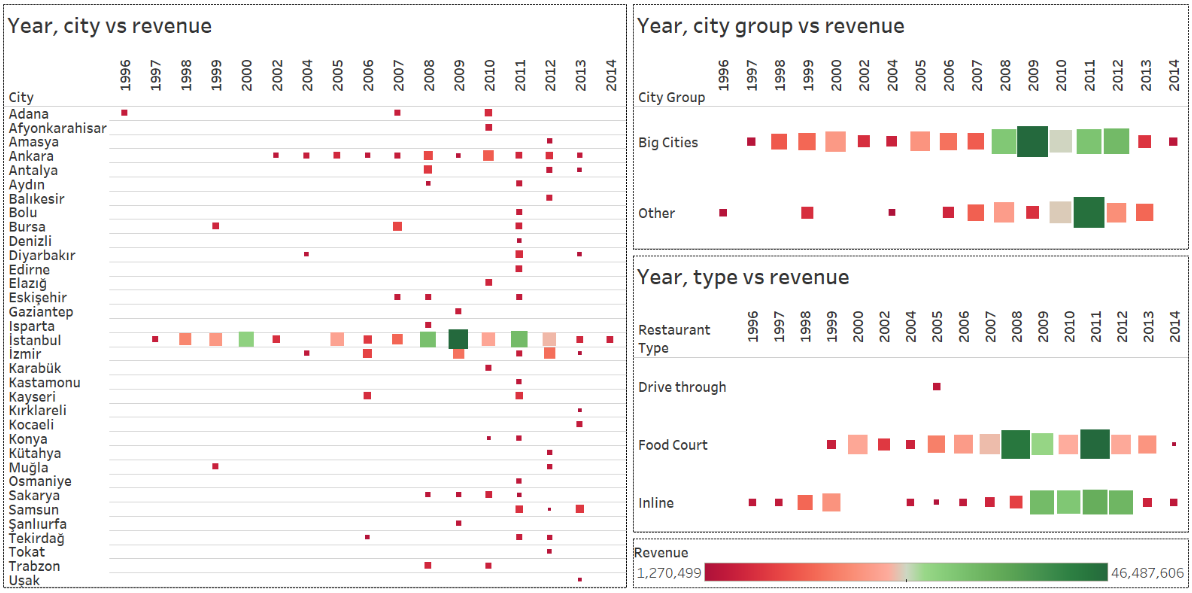


Figure 5: Visualization of relationship between revenue and year with city, city group and restaurant type

The first heat maps above (Figure 5) showing the relationship of revenue and year with city. Insight reveal almost every year Istanbul is the favourite location to open new restaurant and able to generate high revenues. Similarly, Ankara and Izmir are the next preference cities after Istanbul. For relationship of revenue and year with city group discovered few restaurants have opened in Big cities and at same time earned good revenue since 1997 compared to the other category. The final heat map indicates nearly all year, the majority of new restaurants choose food court or inline restaurant type and both capable to yield similar average revenue.

4.2 Data Pre-processing/ Transformation

Orange data mining tools had been used during data cleaning process. Orange tool offers speed of execution and ease of prototyping of new algorithms. Various tasks spanning from data pre-processing to modelling and evaluation are supported as well (Demšar et al., 2004). In the midst of uploading the data, Orange tool have read 'Restaurant ID', 'P-variable' and 'Revenue' as numeric, 'City', 'City group' and 'Restaurant type' as categorical while 'Open date' as meta data. Restaurant ID had been removed from train and test data set as no contribution to the prediction model. The group of restaurant type then been renamed for easy interpretation. 'FC', IL, 'DT' and 'MB' rename to 'Food Court', 'Inline', 'Drive through' and 'Mobile' respectively. Removal of restaurant id had been done in the time of upload the data while edit domain widget had been apply for renaming the group of restaurant type. For 'Open date', Orange data mining required to follow ISO 8601 format to be able recognised as date time categorical. Thus, from excel train dataset, the 'Open date' had been transformed into requirement.

Train datasets were than be checked the missing value by using features statistics and no missing value were found. According to revenue's histogram above (Figure 2) there are outliers existed in the train dataset. Hence, 9 cases had been eliminated by outlier widget and left the total balance of 128 cases to build the model. Based on bar chart of restaurant type shown in Figure 3, there is mobile category in test dataset but not presenting in training dataset. There is high probability difficult for a model to predict when no weights are existed. Thus, 'Mobile' data would combine with 'Drive through' and rename as 'Others' by using edit domain widget since it was the smallest portion of restaurant type before 'Mobile'.

There is not much details information about each list of P-variable. The only information given is the variables contain three categories demographic, real estate data and commercial. In addition, it is unknown if each variable contains a combination of the three categories or are mutually exclusive. Further to my reading in Kaggle's discussion, the host confirm the P-variables were ordinal data but not verified what each P-variables stand for. During uploading the data into orange data mining, P-variables was read as numeric data, thus change to categorical data type was applied. Additionally, based on the observation, majority of the value for some of these features are given zero value and Kaggle community concluded zero value presenting as missing value. Hence, the impute widget were used to give average value to the missing value.

Hyndman and Athanasopoulos (2018) said in their book, it necessary to split the data into two groups as train and test data into 80:20 portions (Kolkova, 2020). While Bera (2021) had list down the validation set that can be applied in machine learning including re-substitution, hold-out and K-fold cross (Bera, 2021). Tentatively, Gogolev and Ozhegov (2019) had applied leave-one-out and 10-fold cross validation for assessing the prediction power of model (Gogolev and Ozhegov, 2019). For this study, 10-fold cross validation had been employed to retrieve the validation of predicting model. In conclusion, 128 samples and 42 features of train dataset would be use for building the model.

5. Results and Discussion

Regression evaluation is aimed at determining the prediction error, hence the result of evaluation metric would be explained by the lower the value of RMSE and MAE, the better the forecasting performance is.

Table 1: Train model evaluation result

Model	Multiple Linear Regression ('000)	Decision Tree regression ('000)	Random Forest regression ('000)	Support vector regression ('000)
RMSE	2,953.92	3,062.26	2,457.00	2,699.26
MAE	2,136.30	2,165.34	1,743.88	1,663.54
Percentage of reduction error	38.3%	41.4%	40.9%	62.3%
Number of observations	128			
Number of predictors	41			

Table 1 presents the results of the four machine learning models with two different evaluation metrics. The table clearly show that the decision tree regression was the least accurate as overall model based on both RMSE and MAE evaluation metrics. According to RMSE result, random forest regression model has the lowest error score but for MAE result presented that the support vector regression was the lowest score among other algorithms. Then I further analyse on percentage of reduction error and found support vector regression was the highest reduction error with the respect to RMSE and MAE.

Thus, as above result, support vector regression gives the lowest error score for restaurant revenue prediction and gives the highest reduction percentage of error in comparison with RMSE and MAE. Therefore, support vector regression outperforms those three algorithms and recommended to be the most suitable forecasting method to apply for this study.

Support vector regression works is by locating the best-fitting line. The hyperplane with the greatest number of points is the best fit line in SVR. SVR's approach implies model estimation based on the training data points closest to the hyperplane. This means that outliers are ignored and the model is automatically trained on observations that are closest to the "average" observation (Gogolev and Ozhegov, 2019). Although SVR uncommonly used, it carries certain advantage. SVR is robust to outliers and decision model can be easily updated. SVR also has excellent generalization capability by technique of maximise the margin and create the largest possible distance between separating the hyperplane (Kotsiantis et al., 2007).

It also works well with high dimensional space and uses scales accordingly the higher it goes and the trade-off between the model complexity and the error can be controlled easily as SVR can deal with both continuous and categorical data by capturing the nonlinear relationships in the data. Hence, why it works well with cases that possessing more dimensions than their sample sizes. As it is a non-parametric technique, no assumptions are needed, and the prediction tends to be very accurate (Mohamed, 2017).

The outperform model then has been applied to test dataset to predict the revenue once the 'restaurant id' been removed and transformed the 'Mobile' and 'Drive through' type into

'Others'. Similar pre-processing techniques have been applied to the test set and follow by the discussion of the insight prediction and finding next.

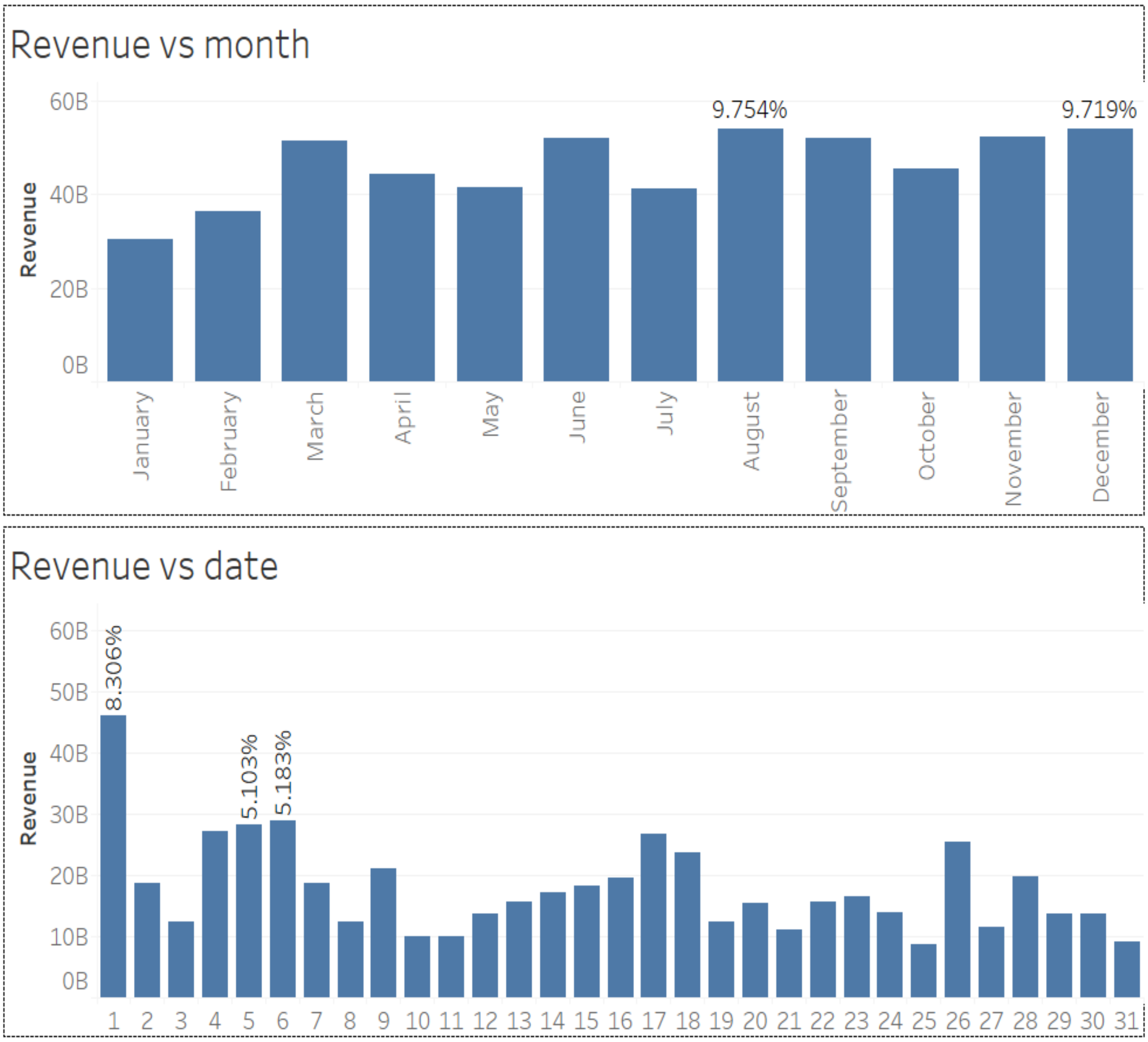


Figure 6: Relationship between revenue and open date for predicted test dataset

Refer to Figure 6, opening new branch in August and December are expected to generate high revenue. This is probably affected by summer and winter season respectively and customer highly tend to buy outside food. Similarly, the best day to start operating the restaurant is on early week of the month specifically on the first, fifth and sixth day of the month. This probably due to customer received salary on early month and allow them to expense more than other weeks of the month. Therefore, TFI should consider opening new branch on August or December during the first week of the month.

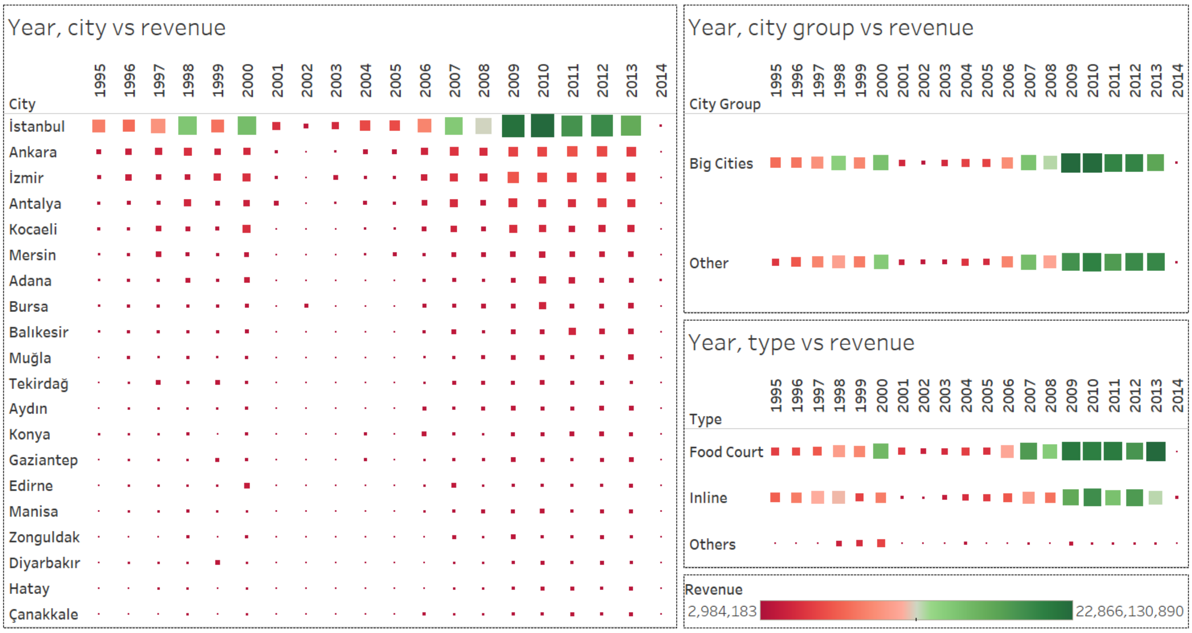


Figure 7: Visualization of relationship between revenue and year with city, city group and restaurant type for predicted test dataset

The first heat maps above observed the relationship of revenue and year with city. Insight reveals Istanbul able to generate high revenues almost every year. Similarly, Ankara and Izmir are the next preference cities to yield high revenue. For relationship of revenue and year with city group shown new branch restaurants opened in Big cities earned good revenue compared to the other category. This is understandably since cities typically have high density of population and have high chance to attract a lot more customer. The final heat map indicates nearly all year, most new restaurants capable to yield high revenue if food court type has been chosen and follow by inline type. This might be due food court able to offer variety type of foods compared to other food service by which customer have narrow range types of food option.

Therefore, finding reveal that TFI should consider opening new branch restaurant on the first week of the month and either on August or December where both can generate good revenue. It is recommended to choose big city in Istanbul for perfect location and food court service is preferable to cater the customer demand in high population and consequently yield high revenue.

6. Conclusions

In this paper, multiple linear regression, decision tree regression, random forest regression and support vector regression (SVR) algorithms had been evaluated to identify which model can effectively predict restaurant revenue for determining the potential location of new restaurant branch. The location evaluation was conducted based on existing branch's revenue, characteristic of restaurant and city attribute. The results of evaluation suggested that support vector regression gives the lowest prediction error and summaries to be superior method in predicting restaurant's revenue. Support vector regression than been applied into test dataset for revenue prediction. The insight reveal opening new branch restaurant on first week of month, either on August or December, located in big city particularly in Istanbul and operate food court service might yield high revenue to the company.

The limitation of this study includes unpresented data in train dataset for evaluating in test data. In test dataset, there are 29 locations that are not presented in the training data. This may generate inaccurate forecasting during the revenue prediction in test dataset. However, I believe it will yield a small prediction error. Moreover, it is challenging to build robust model when the train data consists of 137 data only and need to predict 100,000 test datasets which 729 times bigger. Besides, each of the 37 P-variables has been provided without the details and meaning of each one. The only information offered is demographic, real estate, and commercial data. Thus, the study would have no knowledge if there were redundant variable or not. Therefore, the TFI data that the study selected was inconvenient towards the perfect information. Future study should consider data that will have favourable circumstances for revealing data information.

In addition, the data were collected for the whole TFI's restaurant based on annual revenue including newly open restaurant as well. The presented revenue may not be fair for the newly launch branch restaurant which open less than a year. Thus, the prediction will be more powerful if monthly, weekly or daily revenue can be collected. Furthermore, data were taken from TFI's restaurants, thus the study were limited to food and services industry, thus the obtained result might differ or if applicable for others business industry. However, interested business owner may use this study as a ground to achieve same objective.

For future research, the paper can extend the data wrangling in two ways. It is possible to consider applying one-hot encoding into the object type data. It is a technique of converting data to prepare for an algorithm and achieve a better forecast. One-hot encoding works by converting each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. However, one-hot encoding has the drawback of being longer to process especially when the data contains a lot of categorical value. Next apply principal component analysis for reducing the number of correlated predictors and choose the important predictor that contributes the highest prediction value (Gogolev and Ozhegov, 2019).

In addition, the study should consider a variety of machine learning techniques. Although the results of this study revealed that support vector regression possessed the lowest level of error, it does not guarantee when different algorithms are evaluated, they would produce the same level of error (Hu et al., 2004). Basically, the data provided only take into consideration of location information. This makes it universally relevant, but also restricts the quality of the prediction that can be achieved. In future work, it is possible to add the data source by including information regarding special events and holidays in the study (Posch et al., 2021). By having many relevant predictor variables, it is feasible to confidently discover the best model for predicting restaurant revenue.

In conclusion, this project has presented some important aspects of knowledge of restaurant revenue prediction. This contribution can be use as ground to others business owner who intend to address same objective. The study also presented some limitations and recommendations identified for future studies.

References

Ahmad, I. S., Bakar, A. A., Yaakub, M. R., & Muhammad, S. H. (2020). A survey on machine learning techniques in movie revenue prediction. *SN Computer Science*, *1*(4), 1–14. https://doi.org/10.1007/S42979-020-00249-1

Ariffin, F. F. (2020). *Zara tak putus asa, buka restoran baharu - Berita Harian*. https://www.bharian.com.my/hiburan/selebriti/2020/06/704960/zara-tak-putus-asa-buka-restoran-baharu

Bera, S. (2021). An application of operational analytics: For predicting sales revenue of restaurant. In *Studies in Computational Intelligence*, *907*, 209–235. https://doi.org/10.1007/978-3-030-50641-4_13

Chang, X. & Li, J. (2019). Business performance prediction in location-based social commerce. *Expert Systems with Applications*, *126*, 112–123. https://doi.org/10.1016/j.eswa.2019.01.086

Chen, C.-Y., Lee, W.-I., Kuo, H.-M., Chen, C.-W., & Chen, K.-H. (2010). The study of a forecasting sales model for fresh food. *Expert Systems with Applications*, *37*(12), 7696–7702. https://doi.org/10.1016/j.eswa.2010.04.072

Demšar, J., Zupan, B., Leban, G., & Curk, T. (2004). Orange: From Experimental Machine Learning to Interactive Data Mining. In *European conference on principles of data mining and knowledge discovery* (pp. 537-539). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30116-5_58

Department of Statistic Malaysia. (2020). https://www.dosm.gov.my/v1/index.php?r=column/coneandmenu_id=V3R2ZnB6dE0xU1NDRXNKSHUvdmhkQT09

Gogolev, S. & Ozhegov, E. M. (2019). Comparison of machine learning algorithms in restaurant revenue prediction. *Communications in Computer and Information Science*, *1086CCIS*, 27–36.

Hájek, P. & Olej, V. (2010). Municipal revenue prediction by ensembles of neural networks and support vector machines. *WSEAS Transactions on Computers*, *9*(11), 1255–1264.

Hu, C., Chen, M., & McCain, S.-L. C. (2004). Forecasting in short-term planning and management for a casino buffet restaurant. *Journal of Travel and Tourism Marketing*, *16*(2–3), 79–98. https://doi.org/10.1300/J073v16n02_07

Kolkova, A. (2020). The application of forecasting sales of services to increase business competitiveness. *Journal of Competitiveness*, *12*(2), 90–105. https://doi.org/10.7441/joc.2020.02.06

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3-24.

Mackinlay, J., Hanrahan, P., & Stolte, C. (2007). Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1137–1144. https://doi.org/10.1109/TVCG.2007.70594

McDonald's: number of restaurants worldwide. Statista. (2021). https://www.statista.com/statistics/219454/mcdonalds-restaurants-worldwide/

Mohamed, A. (2017). Comparative study of four supervised machine learning techniques for classification. *Academia.Edu*, *7*(2). https://www.academia.edu/download/54482697/2.pdf

National Restaurant Association. (2019). *Restaurant industry added nearly 14K locations in 2018. NRA. (n.d.).* Retrieved from https://restaurant.org/education-and-resources/resource-library/restaurant-industry-added-nearly-14k-locations-in-2018/

Othman, K. (2019). *Hanya bertahan 3 bulan, Ning Baizura tutup restoran – Hiburan.* m*Star*. https://www.mstar.com.my/spotlight/hiburan/2019/03/30/ning-tutup-kedai

Posch, K., Truden, C., Hungerländer, P., & Pilz, J. (2021). A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2021.06.001

Restaurant Revenue Prediction. Kaggle. (2015). https://www.kaggle.com/c/restaurant-revenue-prediction

Reynolds, D., Rahman, I., & Balinbin, W. (2013). Econometric modeling of the U.S. restaurant industry. *International Journal of Hospitality Management*, *34*(1), 317–323. https://doi.org/10.1016/J.IJHM.2013.04.003

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, *79*(2), 679–683. https://doi.org/10.1016/j.procir.2019.02.042