Compare Linear regression, Decision Tree Regressor, and Random Forest Regressor based on python, a restaurant company on Kaggle as a case

Weicun Zhang[1, *]

[1]School of Art and Science, Rutgers University, New Jersey, USA

*Corresponding author: wz293@rutgers.edu

Abstract. The fastest-developed computer language around the world is python because it is easy to use. Most of the linear regression methods currently used for traffic prediction cannot explain the accuracy differences between methods. In this study, the idea of constructing a regression model, decision tree regressor, and random forest regressor is described by analyzing the sales volume, region, and time of a take-out company, and the accuracy difference between the three different models with the same data is compared. It is concluded that the random forest regression volume has the highest accuracy among the three, and the results are obtained by bringing them into the original data for experiments. Although the linear regression model improves the scientific and rational nature of logistics forecasting. It provides a decision basis for enterprises to forecast sales volume. However, if only the linear regression model is used for forecasting, it is still not accurate enough. It needs to be combined with other methods for experiments to further improve accuracy, scientificity, and rationality. In the paper, the basic knowledge of statistics is used to create three different linear regressions through python and find which one is the most accurate. R2 score, MSE, and RMSE are explained. Based on the study, when firms do the prediction of their business or any other projects, if the data set have the linear regression relationship, they could choose the Random Forest through python directly as it is the most precise among them.

Keywords: Logistics, forecast, python, regressor.

1. Introduction:

Although, right now, python is the fastest-developed computer language all over the world, based on its practicability, efficient learning curve, and high-quality packages of data, R language is used much more often than it in statistics [1]. Python could also deal with some problems and give predictions based on statistical formulas.

Forecasting becomes even more important as companies expand globally and the supply chain of products and services lengthens. This is important for planning and managing activities [2]. In recent years, data forecasting has evolved very rapidly, which largely provides a scientific reference for the transformation of corporate business models [3]. As an important part of logistics system planning, the prediction accuracy of logistics flows forecasting directly affects the science and effectiveness of logistics system analysis, modeling, planning, and layout [4]. Therefore, it is important to choose more accurate forecasting methods to improve the quality of logistics planning and operational efficiency. Therefore, choosing a more accurate forecasting methodology is critical to improving the quality and operational efficiency of logistics planning. Accurate forecasts help companies and supply chains prepare for long-term changes in short- and long-term market conditions and improve operational performance. Forecast-based decisions can lead to operational errors if forecasting accuracy declines. Similar relationships have been demonstrated in logistics and business research to help assess the impact of forecast accuracy on replenishment, supplier management, inventory management, joint planning forecasts, and service component management. Such studies also assess forecasts as "meaningful terms to managers" through commonly used statistical measures that help quantify the impact of forecast performance, including inventory investment and service achievement. In a highly competitive marketplace, SMEs that make the most accurate judgments about the marketplace can plan strategic growth plans based on actual conditions, continuously improve management capabilities, and cultivate competitiveness [5].

There are many methods currently used for logistics forecasting, such as trend extrapolation, gray forecasting, neural network forecasting, etc [6]. These methods essentially establish a fitting model of the original data to maximize the fitting accuracy and perform forecasting analysis accordingly. However, no single prediction method can be applied to all problems to reduce operational errors and improve prediction accuracy. Linear regression models are often utilized in different fields. Depending on the situation, the different models' performance is excellent. When applying the Random Forest Regressor to the field of renewable energy, the average prediction accuracy was about 93% [7]. Overall, the Random Forest Regression (RFR) model consistently outperforms other models and is well-suited for real-time situational awareness deployments that identify the location and duration of failures while addressing missing data [8].

This study uses and compares R2 scores, MSE scores, and RMSE from three different regression models in Python to determine which model can predict sales. Three different statistical values were obtained by testing the same dataset and comparing the accuracy of different models in the same example to select the model with the highest accuracy. Finally, the most accurate model is added to the original data for prediction and the results are obtained. By comparing different models, the accuracy of business forecasting can be further improved, thus reducing errors and helping companies to make planned strategic plans for growth according to the actual situation. This study uses python to simply predict enterprise logistics, filling the gap of not using python in the logistics forecasting process, and comparing the differences in predictions of different linear regression models on the same data set. Random Forest Regressor is the most precise among three different linear regression models based on the food company.

The rest of the paper is organized as follows. Section 2 contains the process of dealing with the data set. It analyzes how to do the different model work through the same test set. The results are discussed in Section 3. Section 4 concludes the paper and provides some advice to improve the linear regression or how to improve the precision for the company to forecast.

2. Methodology

2.1 Data

The study utilizes the data set from Kaggle. It is a meal delivery company that operates in multiple cities. This is a food delivery company operating in several cities. These cities have various distribution centers that deliver meal orders to customers. The customer wants these centers to be able to forecast demand for the next few weeks and plan their raw material inventory accordingly. Since most raw material replenishment is done every week and raw materials are fragile, procurement planning is of utmost importance. Second, accurate demand forecasting is another very useful aspect. Testing is based on the test sets provided by the firm.

2.2 Methodology

2.2.1 Import and train data

Four excel sheets and another sheet that combines the four features are provided. The four tables of the train, test, meal info, and fulfillment center info are utilized and imported in python. At the same time, the head loop is used to detect whether there is an error. It is used to print a concise summary of a Data Frame. From the table, the range could be found, which is 456548 entries from 0 to 456548, and 9 data columns.

For numeric data, make a histogram to understand the distribution and Corrplot. For categorical data, make a histogram to understand class balance. Fig.1 describes the sales amount of different weeks from excel. Fig.2 displays the number of sales when the checkout prices are different.

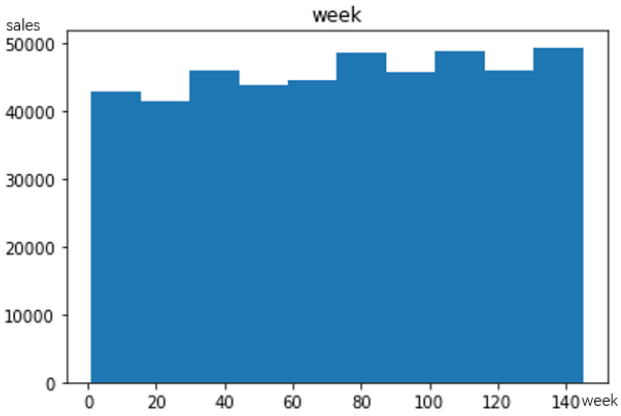

Fig. 1 Week with sales amount

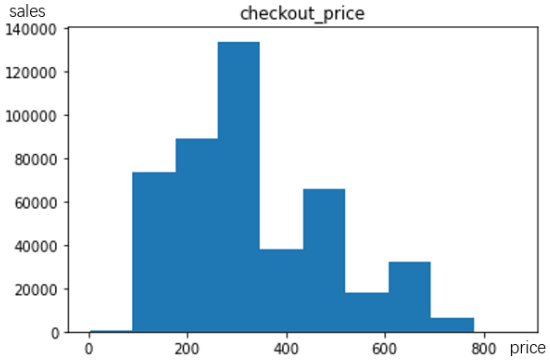

Fig. 2 Checkout price with sales amount

Then a heatmap is to determine the relationship between the two. Correlation refers to a statistical relationship between two random variables or bivariate data, whether causally related or not. In the broadest sense, correlation refers to any statistical association but usually refers to the degree to which a pair of variables are linearly related. Fig.3 is a heatmap of weeks and checkout price, which is to check whether the two vectors have a linear relationship. It describes the linear relationship between weeks and checkout price. Then Fig.4 is taken all the vectors in the data set and creates a heatmap.

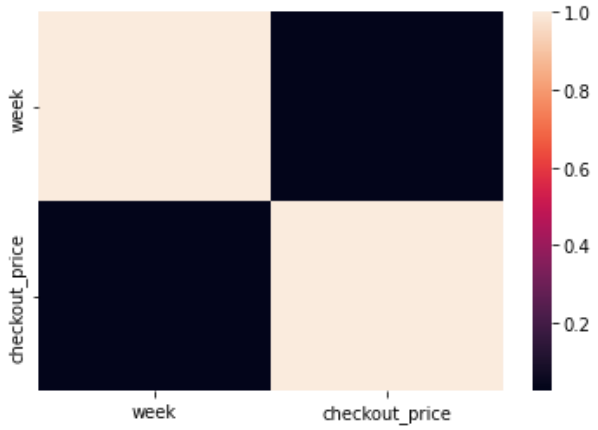

Fig. 3 features between a week and check-out price

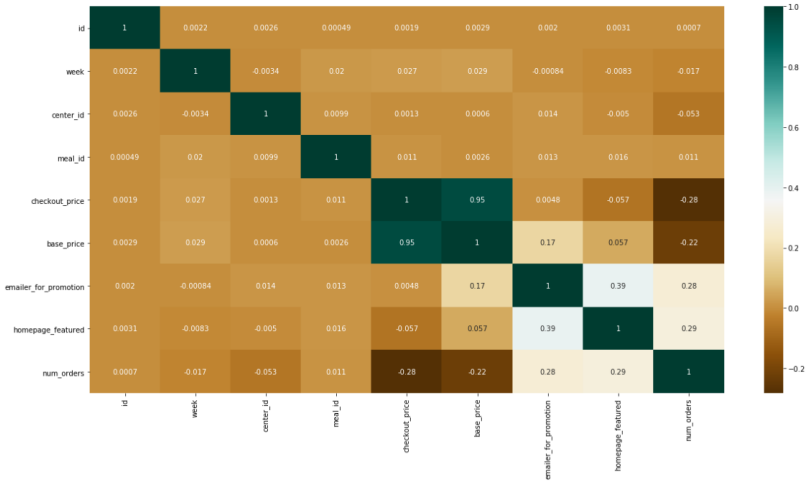

Fig. 4 Features of nine elements

2.2.2 Data normalization

A boxplot is used for data normalization. the Q1 and Q3 are 25% and 75% percentile. IQR (interquartile range) is come out and uses the formula. Then the Upper range and Lower range from the formula are found. In Fig.5, the Q1 is between 20 and 40, and Q3 is between 100 and 120. Then the numbers are plugged into the formula above. Fig.6 is a boxplot about the checkout price, which is similar to weeks.

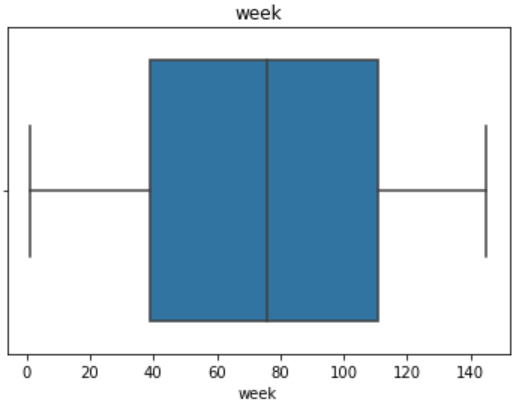

Fig. 5 Boxplot of week

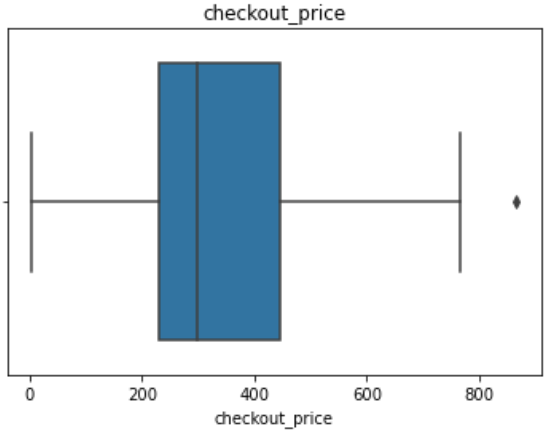

Fig. 6 Boxplot of checkout price

2.2.3 future selection

The Pair plot figures out the relation between features and labels, based on the data at the beginning. From the pair plot of the seven elements of Fig.7, there is a linear relationship between the base price and checkout price. For other elements, there is no obvious relationship. Simple linear regression models could be used to test.

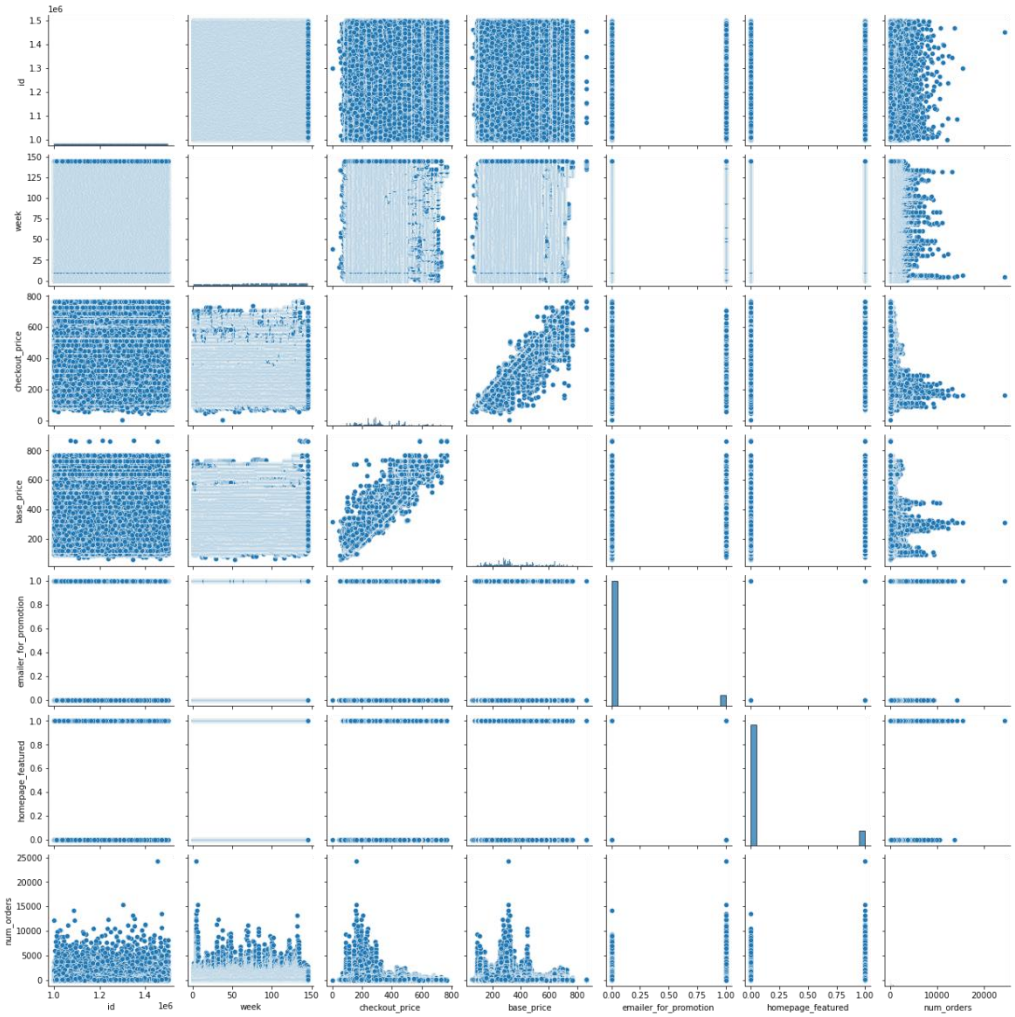

Fig. 7 Pair plot of seven elements

2.2.4 model building

According to the heat map in the first part and the pair plot in the third part, build a linear regression model in python to make predictions. To make the predicted value more accurate, three models of linear regression are utilized, and calculate their R2 score, MSE score, and RMSE for comparison.

SSR (Regression sum of squares): It is the error between the estimated value and the mean value. SSR reflects the sum of squared deviations of the degree of correlation between the independent and the dependent variable.

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad (1)$$

SSE (Residual Sum of Squares): Error between the estimated and true values, reflecting the fitness of the model

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

SST (Sum of Squares of Total Deviations): This is the Sum of SSR and SSE

$$SST = SSR + SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (3)$$

R2-score: A coefficient of determination that reflects the proportion of all changes in the dependent variable

$$R2 = 1 - SSE/SST \quad (4)$$

The MSE (mean squared error) score is a measure that reflects the degree of variation between the estimated quantity and the estimated quantity. When the predicted value matches the true value exactly, it is 0, which is a perfect model; the larger the error, the larger the value. In summary, the smaller the value, the more accurate the machine learning network model is, and vice versa, the worse it is.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2 \quad (5)$$

RMSE (Root Mean Square Error): RMSE is the square root of MSE. In actual measurements, the number of observations n is always limited, and only the most reliable value can be used for the true value. In the same way as MSE, the smaller the difference between the predicted value and the real value, the higher the model accuracy.; on the contrary, the lower it is.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2} \quad (6)$$

The linear regression model is

$$y_t = x_t'\beta_t + u_t, \quad t = 1, \dots, T, \quad (7)$$

y is the dependent variable, x is the vector of observed values of the independent variable, b is the vector of unknown regression coefficients, and u is the unobservable disturbance term. [9].

Decision tree regression is a tree-based structure used to predict numerical outcomes for a dependent variable [10]. Decision tree is a nonparametric supervised learning method for classification and regression. Decision tree regression was chosen for this study because, unlike traditional decision trees, it predicts numerical outcomes for the dependent variable.

Random Forests is a machine-learning algorithm that uses a collection of decision trees. This algorithm combines bagging method and the random subspace method. It is usually used for classification, regression, and clustering [11]. The main idea is to use a large collection of decision trees. Although the classification results for each of the decision trees are of very low quality, the large number of them makes the results accurate enough.

3. Results

Table 1. Comparison results between models

Model/score	Linear Regression	Decision Tree Regression	Random Forest Regressor
R2 score	0.22430653751828467	0.2615191471545507	0.2596071920579188
MSE score	111857.88773504822	110773.22350524069	109570.40278549175
RMSE score	334.45162241353864	332.8261160204239	331.0142032987282

Based on the values of RMSE and MSE it is easy to find out that linear regression is used to forecast the restaurant companies. Random Forest Regressor is the most accurate of the three because

it has the lowest value. Although the random forest regressor has the lowest value among the three, it still has 331 if people only look at the RMSE values, which means that it still has a large error, and it is easy to see from the pair plot that the linear relationship of this company is not very strong. Therefore, there is a need to use some other methods to solve the problem. When using a random forest regressor, there is very little data to prepare. Other techniques typically require data normalization, the creation of dummy variables, and the deletion of blank values. Note, however, this module does not support missing values. Statistical tests can be used to validate the model. Moreover, it works well even if the actual model that generates the data violates its assumptions in some way.

However, just because a random forest regressor is the most accurate for this firm does not mean it is the most accurate for all others. The linear regression model is also flawed. Random forests are not as good at solving regression problems as a classification because they do not produce successive outputs. When a random forest makes predictions beyond the training set data, it may lead to overfitting when modeling noisy data. As can be seen from the table, both the MSE and RMSE scores are too high. However, this does not mean that the three different models are inaccurate. This is because both the MSE and RMSE scores are sensitive to extreme data. Therefore, to get more accurate predictions, it is required to define a range and discard some extreme values on the graph. In this case, at the beginning of the process, the data of seven elements should be cleaned to fix the problem.

4. Conclusions

By comparing the three different linear regression models, obtained the R2 score, MSE score, and RMSE score, and by comparing the three scores it concluded that the random forest regressor is the most accurate of the three. Under the condition of the same training set, the linear regression results are easy to understand and computationally uncomplicated. One of the advantages of a decision tree regressor is that it does not require any feature transformation if people are dealing with nonlinear data. The decision tree does not stop the growth of iteratively splitting the nodes until the leaves are pure or the top condition is satisfied. Random Forest Regressor Training is highly parallelizable, which is advantageous for the speed of training large samples in the era of Big Data. The ability to randomly select decision tree nodes and partition features allows for efficient training of models even when the dimensionality of the sample features is high. Therefore, when comparing these three regression models as linear regression models, random forest regression is more accurate.

This study found that linear regression is feasible for company prediction and random forest regressor is the most accurate model among the three by python with different linear regression models. However, the RMSE value is still too high for the company. The reason that the score is 331 is that some extreme data and large numbers are inside the model. RMSE is easily impacted by these large numbers because it is the root of MSE. Using a random forest regressor to predict the company can help to judge the company, but to reduce the error in company decisions and further improve the accuracy of company prediction, add other models in python to help the company judgment. To improve the model in the paper, before the model is created and tested, data cleaning is necessary. After cleaning some extreme data, RMSE would not be overfitting and A better result would be found.

In this paper, there is a linear relationship among the data of the food company. However, for the fitting of some nonlinear data, it is possible to use other models. For instance, to increase the stability of the linear regression model, further optimizing the model could be used. An example is a new test for the constant regression coefficient of a linear model. The test does not require any knowledge of the likelihood of the change point. Obtain the limiting null distribution of the test statistic and show the non-trivial ability of the test for many local cases.

Moreover, after the company tried the random forest regressor, other methods like trend extrapolation, gray forecasting, and neural network forecasting, could be used to make sure the accuracy of forecasting.

References

[1] Pingouin: statistics in Python by Raphael Vallat1, journal of open source software DOI: 10.21105/joss.01026

[2] Wagner, S. M. Innovation management in the German transportation industry. Journal of Business Logistics, 2008, 29(2): 215.

[3] Zhang, L. Regional Object Flow Forecasting Method Simulation in the Internet of Things Environment. Computer Simulation, 2018, 35(8): 410-414.

[4] Smith, C. D., Mentzer, J. T. User influence on the relationship between forecast accuracy, application, and logistics performance. Journal of business logistics, 2010, 31(1): 159.

[5] Hui, Y., Bo, W., Cui, S., Yan, R. C. Logistics Forecasting Based on Support Vector Machine: A Case Study of Chengdu.Transportation Standardization, 2014, 42(15).

[6] Hu, B. Y., Wang, X. K., Chen, Y. Capacity prediction of logistics park based on double improved gray Markov model. Journal of Dalian Maritime University, 2011, 37(1): 91-94.

[7] Khalyasmaa, A. Prediction of Solar Power Generation Based on Random Forest Regressor Model. International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), 2019.

[8] Mrabet, Z. E. Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems. the Special Issue Cybersecurity and Privacy-Preserving in Modern Smart GridSensors. 2022, 22(2): 458.

[9] Ploberger, W., Kramer, W., Kontrus, K. A New test For Structural stability in the linear regression model. Journal of Economics, 40(1989): 307-318.

[10] Rathore, S. S., Kumar, S. A Decision Tree Regression based Approach for the Number of Software Faults Prediction. ACM SIGSOFT Software Engineering Notes, 2016, 41(1).

[11] Khalyasmaa, A. Prediction of Solar Power Generation Based on Random Forest Regressor Model. International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), 2019.