Project Report: MBA664A
Under Prof. Vipin B.
Members: Deepankur Kansal, Ishan Gupta and Nikunj Jain.
(Group 7 : 3 members only)

Introduction

In recent years, the rise of the Internet of things (IoT) as an emerging technology has been unbelievable, more companies are moving towards the adoption of these technologies and many IoT sensors are being deployed to share information in real-time which leads to the generation of a huge amount of data. This data when used correctly, will be very helpful to the company to discover hidden patterns for better decision making in the future. For example, with the DataCo company, dataset customer segmentation analysis was performed in this project which helps the company to better understand its customers and target them to increase customer responsiveness and the company's revenue. With a lot of options available to analyze data, it is very difficult to decide which method and machine learning model to use since the performance of the model vary on the parameters available in the data.

With the growth of machine learning, there have been numerous comparison studies that compare the performance of neural networks with traditional linear techniques for forecasting. For example, author Carbonneau et al. (2007) in their research work compared various traditional forecasting time-series like moving average, linear regression with recurrent neural networks and support vector machines and concluded that recurrent neural networks performed best. Hill et al. (1996) have also considered the M-competition data and have compared neural networks and traditional methods. Vakili et al. (2020) evaluated the performance of 11 popular machine and deep learning algorithms for classification tasks using six IoT-related datasets and concluded that Random Forests performed better than other machine learning models, while among deep learning models, ANN and CNN achieved more interesting results. Some other authors like Ahmed et al. (2010) did a study comparing different regression models and concluded that the MLP model and Gaussian process models are the best two models for regression type data. But no study that compared both Classification type ML models and Regression type ML models against the Neural Network models with the same dataset was found.

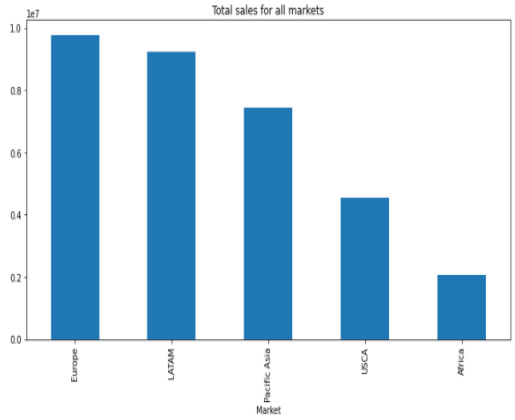
This project aims to compare 9 popular machine learning classifiers and 7 regressors type machine learning models and measure their performance against neural network models to find out which machine learning model performs better. Since the dataset used is related to the supply chain important parameters are identified and the machine learning models are trained with the dataset for detection of fraud transactions, late delivery of orders, sales revenue and quantity of products that customer orders. The machine learning classifiers used in this project are Logistic Regression, Linear Discriminant Analysis, Gaussian Naive Bayes, Support Vector Machines, k - Nearest Neighbors, Random Forest classification, Extra Trees classification, Extreme Gradient Boosting, Decision Tree classification for fraud detection and to predict late delivery on the basis accuracy, recall score and F1 score. The regression models used are Lasso, Ridge, Light Gradient boosting, Random Forest regression, Extreme Gradient Boosting regression, Decision Tree Regression, and Linear Regression to predict sales and quantity of the products required which are compared with mean absolute error (MAE) and root mean square error (RMSE).

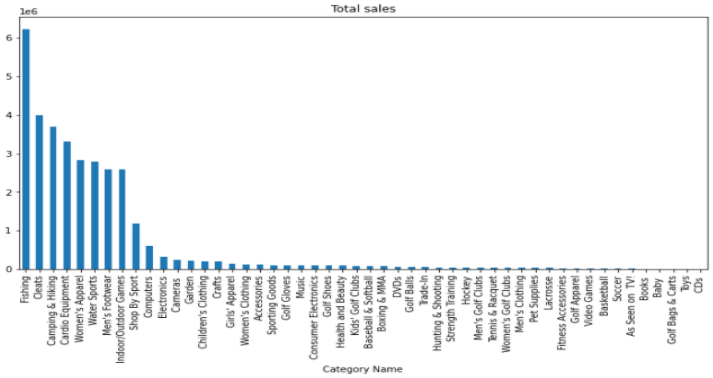
Dataset

A DataSet of Supply Chains used by the company DataCo Global was used for the analysis. The dataset consists of around 180,000 transactions from supply chains used by DataCo Global for 3 years. Link . The data being used has 180519 records and 53 fields. The data consists of some missing values from Customer Lname, Product Description, Order Zipcode and, Customer Zipcode which should be removed or replaced before proceeding with the analysis.

Data Visualization

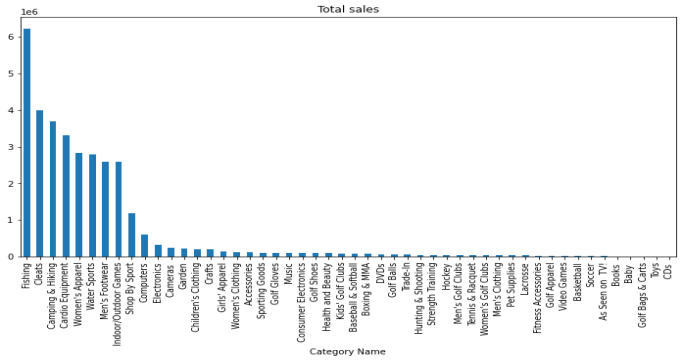
Which region has maximum sales? We plot out the sales from different regions and by correlation plots we can observe that product price has a high correlation with Sales, Order Item Total.

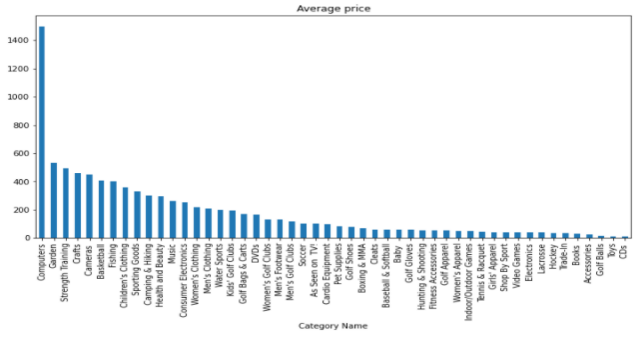




It could be seen from the graph that the European market has the most number of sales whereas Africa has the least. In these markets, western European regions and central America recorded the highest sales.

Which category of products has the highest sales? We plot out the total sales, average sales and average price of each product category.



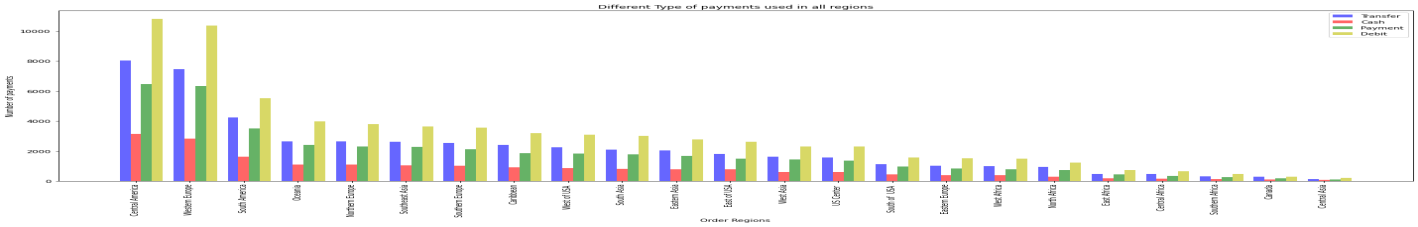


As we can see from fig 1 that the fishing category had the most sales followed by the Cleats. However, it is surprising to see that the top 7 products with the highest price on average are the most sold products on average with computers having almost 1350 sales despite the price being 1500$. Since correlation was high between Price and Sales it will be interesting to see how price is impacting the sales for all the products to see the trend.

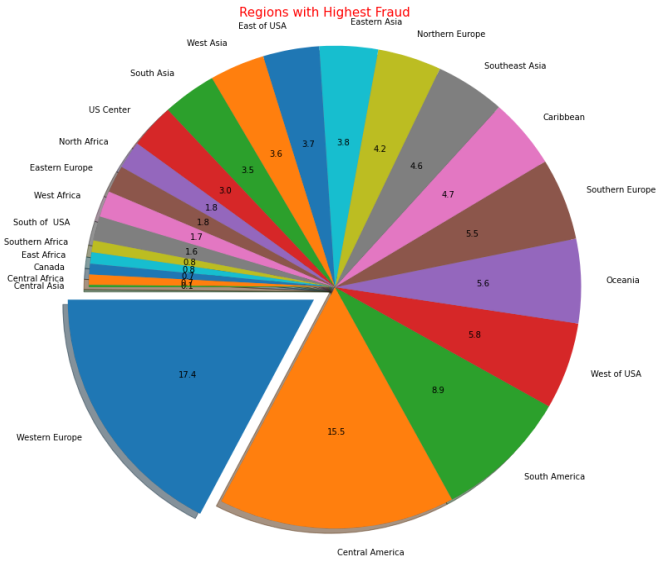
Which quarter recorded the highest sales?
It can be found by dividing order time into years, months, weekdays, hours to better observe the trend.

The most number of orders came in October followed by November, and orders for all other months are consistent. The highest number of orders are placed by customers in 2017. Saturday recorded the highest number of average sales and Wednesday with the least number of sales. The average sales are consistent throughout the day irrespective of time with an std of 3.

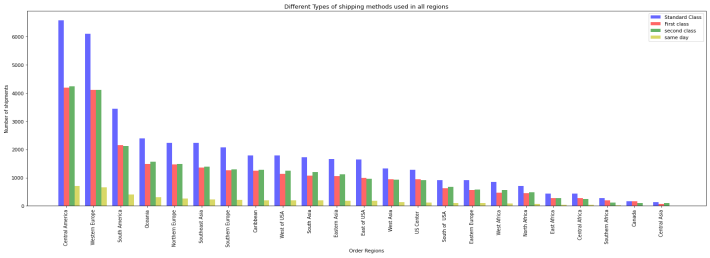
Which region and what product is being suspected of the fraud the most?

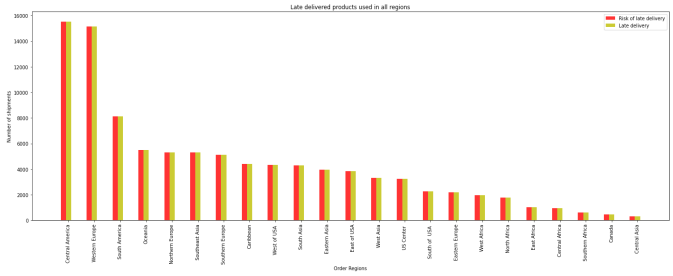


The category which showed the most losses were cleats while region-wise it is observed that Central America and Western Europe make the highest losses. The total loss is 3.9 Million dollars.



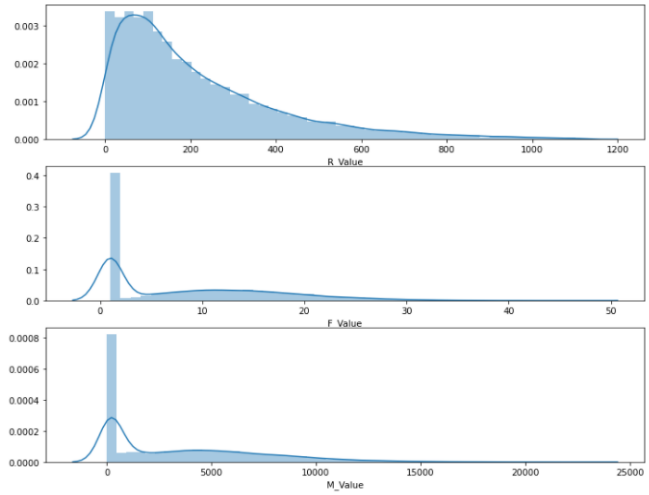
It can be observed that the highest number of suspected fraud orders are from Western Europe which is approximately 17.4% of total orders followed by Central America with 15.5%. It can be seen that orders with Cleats department are getting delayed the most followed by Men's Footwear. The total amount was almost 102k which is a very huge amount. Since Mary was using a different address every time when placing orders, a new customer id was issued each time which makes it difficult to identify the customer and ban them. All these parameters should be taken into consideration to improve the fraud detection algorithm so fraud can be identified more accurately.

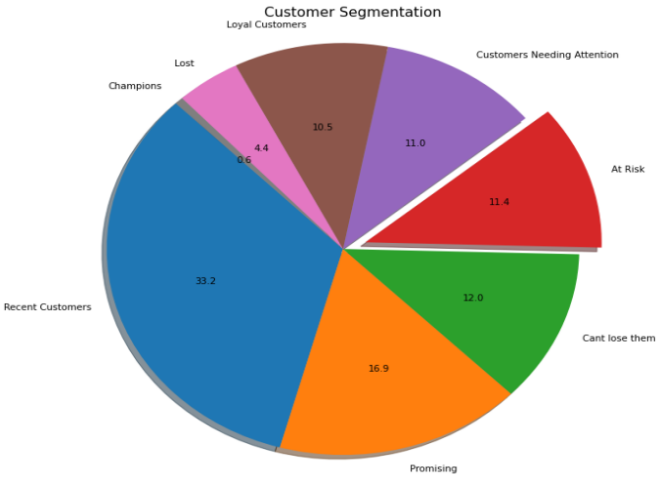




Customer Segmentation

One strategy for a supply chain organisation to improve the number of customers and earnings is to understand customer demands and target specific clusters of customers based on those needs. Customers' purchase histories are already available in the dataset, so RFM analysis can be used for consumer segmentation. Despite the fact that there are other ways for customer segmentation, RFM analysis is popular because it uses numerical numbers to display client recency, frequency, and monetary values, and the output findings are simple to interpret.





In the above plots, R_Value(Recency) indicates how much time elapsed since a customer's last order, F_Value (Frequency) indicates how many times a customer ordered, M_Value (Monetary Value) tells us how much a customer has spent purchasing items. The R_Value should be low because it indicates recent customer activity and F_value, M_Value should be high since they indicate the frequency and the total value of purchase. To make it easier for segmentation individual R, F, M scores were added together. Since total clients are separated into nine divisions, it can be seen that 11.4 per cent of consumers are at risk of losing their business, and 11% of customers require immediate attention or they would be lost.

Data Modelling

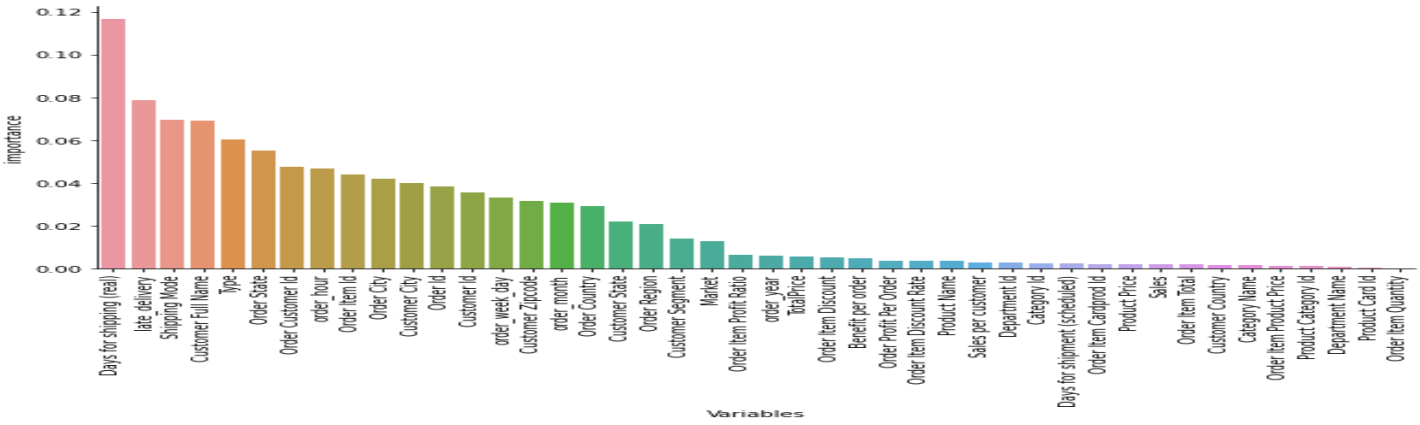
To measure the performance of different models the machine learning models are trained to detect fraud, late delivery for classification type. And sales, order quantity is predicted for regression type models. Two new columns are created for orders with suspected fraud and late delivery making them into binary classification, which in turn helps to measure the performance of different models better. Now to measure machine models accurately all the columns with repeated values are dropped like the late_delivery_risk column because it is known all the products with late delivery risk are delivered late. And Order Status column because a new column for fraud detection is created there is a chance machine learning model might take values directly from these columns to predict the output.

Considering the F1 score it is clear that the Decision Tree classifier is performing better for classification type with an F1 score of almost 80% for fraud detection and 99.42% for late delivery. Surprisingly, all the models except the Russian model predicted the late delivery of orders with almost 98% accuracy. Just to make sure that the model is predicting correctly the model is cross-validated and the results are compared with the accuracy of the model.

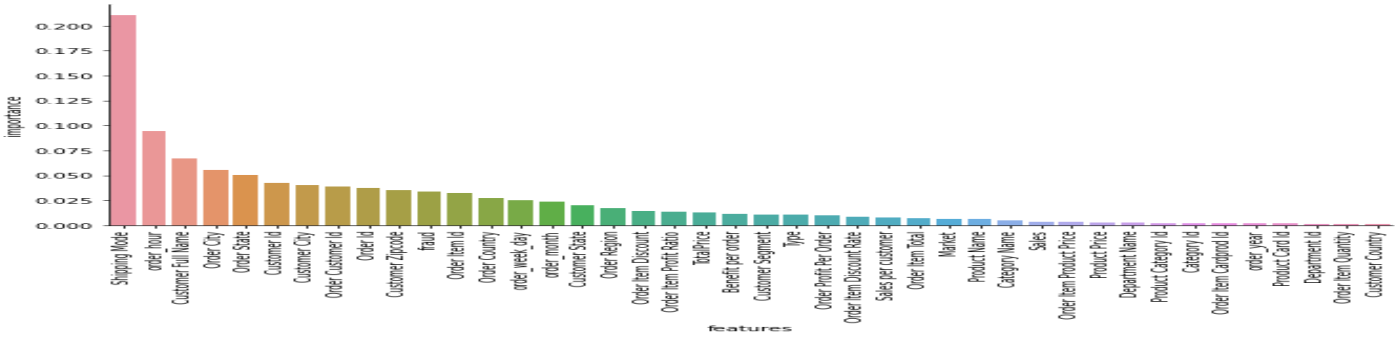
	Classification Model	Accuracy Score for Fraud Detection	Recall Score for Fraud Detection	F1 Score for Fraud Detection	Accuracy Score for Late Delivery	Recall Score for Late Delivery	F1 Score for Late Delivery
0	Logistic	97.80	59.40	31.22	98.84	97.94	98.96
1	Gausian Naive bayes	87.84	16.23	27.92	57.27	56.20	71.95
2	Support Vector Machines	97.75	56.89	28.42	98.84	97.94	98.96
3	K nearest Neighbour	97.36	41.90	35.67	80.82	83.45	82.26
4	Linear Discriminant Analysis	97.88	56.57	49.20	98.37	97.68	98.52
5	Random Forest	98.48	93.18	54.57	98.60	97.52	98.74
6	Extra trees	98.61	98.88	58.60	99.17	98.51	99.25
7	eExtreme gradient boosting	98.93	89.89	73.22	99.24	98.65	99.31
8	Decision tree	99.12	82.53	81.00	99.37	99.44	99.42

The f1 score for the neural network model is 96.48% which is pretty high and better when compared with the decision tree f1 score which was 80.64. But comparing accuracy scores it can be concluded that even machine learning models did pretty good for fraud detection and late delivery prediction.

Feature Importance



Even though fraud detection is not at all related to Days for shipping(real) it is very surprising to see it was given an importance of 0.12. All other important parameters like customer full name, shipping mode, type of payment used are given the importance of 0.7 which helps the company to detect fraud accurately when the same customer is conducting fraud.



This time variables like shipping mode, order city, state are given more importance which helps the company to use different shipping methods to deliver products faster.

	Regression Model	MAE Value for Sales	RMSE Value for Sales	MAE Value for Quantity	RMSE Value for Quantity
0	Lasso	1.5500	2.3300	0.9000	1.030
1	Ridge	0.7500	0.9700	0.3400	0.520
2	Light Gradient Boosting	0.4600	1.6600	0.0010	0.011
3	Random Forest	0.1900	1.7900	0.0001	0.006
4	eXtreme gradient boosting	0.1540	3.1300	0.0005	0.004
5	Decision tree	0.0130	0.9180	3.6900	0.006
6	Linear Regression	0.0005	0.0014	0.3400	0.520

The MAE and RMSE scores for neural network models are 0.007 and 0.022 which are pretty good. But surprisingly, the MAE and RMSE scores were lower for Random Forest and eXtreme Gradient Boosting ML models.

Conclusion

Analyzing DataCo's corporate data, both Western Europe and Central America are the best-selling regions, while companies are also losing most revenue from these regions. Also, delivery is delayed due to most of the fraudulent transactions and orders in both spaces. The company's total revenue remained constant until the third quarter of 2017, with total revenue increasing by 10% quarterly and declining by nearly 65% in the first quarter of 2018. October and November are the highest-selling months of the year as a whole. Most people prefer to pay with a debit card, and all fraudulent transactions are processed by wire transfer. As a result, the company has been scammed by more than 100,000 individual customers, so customers who use wire transfers need to be careful. All orders at risk of delivery delays will be delayed each time. When ordering items in the cleats, men's shoes and women's clothing categories, delivery is almost always delayed. The Neural Network Classifier trained for fraud detection performed best with an f1 value of 0.96. Compared to other machine-learning classification models, the decision tree model was well suited to identify subsequent delivery orders and detect fraudulent transactions with an f1 score of 0.80. While linear regression models are good predictors of sales, both Random Forest and eXtreme Gradient Boosting regression forecasts have lower MAE and RMSE values than neural network models, allowing for more accurate demand forecasts. However, the difference between the MAE and RMSE values of the regressor models of neural networks and these ML models is minimal. Amazingly, the Random Forest and eXtreme Gradient Boosting models are superior to the neural network models

Future Work

For further investigation, we can compare all machine learning models against different datasets to see if the performance of the model improves. In addition, we can improve the performance of these machine learning models by hyperparameter tuning.

References

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econometric Reviews, 29(5-6), 594–621.

bars, P., Smith, J., & Lyon, J. (2020). Python matplotlib multiple bars. Retrieved 17 April 2020, from https://stackoverflow.com/questions/14270391/python-matplotlib-multiple-bars

Building Neural Network using Keras for Classification. (2020). Retrieved 18 April 2020, from https://medium.com/datadriveninvestor/building-neural-network-using-keras-for-classification-3a3656c726c1

Carbonneau, R., Laframboise, K., & Vahidov, R.(2008). Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research, 184(3), 1140-1154.

Constante, Fabian; Silva, Fernando; Pereira, António (2019). DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS, Mendeley Data, v5. Retrieved 25 March 2020, from http://dx.doi.org/10.17632/8gx2fvg2k6.5#file-5046ef5f-6df4-4ee7-9eb8-b33456b0d49e

Explaining Feature Importance by example of a Random Forest. (2020). Retrieved 15 April 2020, from https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e

Ferreira, K.J., Lee, B.H.A., & Simchi-Levi, D.(2016). Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1), 69-88.

Find Your Best Customers with Customer Segmentation in Python. (2020). Retrieved 9 April 2020, from https://towardsdatascience.com/find-your-best-customers-with-customer-segmentation-in-python-61d602f9eee6

Hassan, C.A., Khan M.S., & Shah, M.A.(2018). Comparison of Machine Learning Algorithms in Data classification. 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, United Kingdom, 2018, pp. 1-6.

Martinez, A., Schmuck, C., Pereverzyev Jr, S., Pirker C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. European ournal of Operational Research, 281(3), 588-596.

Resampling time series data with pandas – Ben Alex Keen. (2020). Retrieved 7 April 2020, from https://benalexkeen.com/resampling-time-series-data-with-pandas/

RFM Segmentation | RFM Analysis, Model, Marketing & Software | Optimove. (2020). Retrieved 10 April 2020, from https://www.optimove.com/resources/learning-center/rfm-segmentation

sklearn.linear_model.LinearRegression — scikit-learn 0.22.2 documentation. (2020). Retrieved 14 April 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

trees?, H., & Dixit, H. (2020). HOW TO LABEL the FEATURE IMPORTANCE with forests of trees?. Retrieved 10 April 2020, from https://stackoverflow.com/questions/37877542/how-to-label-the-feature-importance-with-forests-of-trees

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. arXiv preprint arXiv:2001.09636.