Weekly Sales Forecasting

GitHub: https://github.com/srjth19/Weekly_Sales_Forecast.git

Rohan Gupta
20210648
School of Engineering and Computing
Dublin City University
Dublin, Ireland
rohan.gupta5@mail.dcu.ie

Srijith Unni
20211114
School of Engineering and Computing
Dublin City University
Dublin, Ireland
srijith.unni2@mail.dcu.ie

Soumi Mitra
20210300
School of Engineering and Computing
Dublin City University
Dublin, Ireland
soumi.mitra2@mail.dcu.ie

Sushma Suryanarayana Gowda
20211607
School of Engineering and Computing
Dublin City University
Dublin, Ireland
sushma.suryanarayanagowda2@mail.dcu.ie

***Abstract*—Prediction of sales is one of the most important factors to any organization. Through the prediction of sales the organizations can actually plan for the demand throughout the year. These predicted sales not only help the operations of the organization go smoother but also boosts the morale of the workers. In this paper we have predicted the weekly sales by training eight different regression models and chose the best one by measuring their performance using (RMSE, MAE, MSE R^2,) The chosen dataset is an historic dataset of an organization which has the data present in 4 separate files.**

Keywords—Prediction, Error Score, RMSE, MAE, R^2, historic dataset.

I. Introduction

Sales forecasting is the way toward assessing future income by foreseeing the measure of items or administrations a sales unit (which can be an individual salesperson, a sales group, or an organization) will sell in the following week, month, quarter, or year.

The Sales Forecasting for any type of organization is very important in order to make decisions that reflect in the organization's growth, like understanding the demands in prior.. It is almost impossible to forecast perfectly but even a forecast within the 10% of actual results can positively impact an organization's business. The decision makers actually rely on these predictions to plan for their business expansion and to determine how to fuel a company's growth.

Our motivation to work on the sales prediction is to understand what factors affect the increase and decrease of sales and which prediction model suits the best for the prediction of the sales.

The aim is to build a model that will predict the Weekly Sales of any organisation through the historic data which is based on various parameters like Store opening hours, holidays, region based information such as temperature, closing timings and of course the holiday seasons.

A. Dataset

The Dataset in the study is provided by "Walmart" and is publicly available on Kaggle. The dataset is divided into Features, Stores, Test and Train data files. The Test data file is the historical training data, which covers the sales record from 2010-02-05 to 2012-11-01 that contains: Store, Dept, Date, Weekly_Sales & IsHoliday. The training dataset has 45 unique stores and 81 unique departments. The weekly forecast is not provided for all 81 departments for each store, the number of entries for each store differs from one another. We also don't know what each of the departments stands for. There are negative weekly forecast values which also needs some investigation.

The test dataset has all the similar parameters to the training dataset except the "Weekly Sales" column. The Store dataset has the following columns: Store, Size & the Type. The features data file is the one that actually contains the features of the dataset it covers: Store, Dept, Date, WeeklySales, IsHoliday, Unemployment, etc.

B. Data Pre-Processing

Pre-processing tasks such as combining datasets, transforming, feature-engineering and handling missing values are performed in Python.

II. Related Work

Companies always search for better data mining techniques to gain higher revenue and maintain the data which is critical. They face various issues to identify the most accurate data mining technique and prediction strategy which will be the most effective. Various data mining techniques, machine learning algorithms and time series model analysis are required for this. In this study, we have referred to several previous works to get the ultimate output.

In 2018, Cheriyan et. al. [1] published a paper where three machine learning models are used for sales prediction. Those models are: Generalized Linear Model, Decision Tree Model and Gradient Boosted Trees. The evaluation metric used here are: Accuracy Rate (%), Error Rate, Precision, Recall and Kappa. Here the Gradient Boosted Tree comes out to be the pioneer one as it achieved 98% accuracy rate.

In 2019, a paper by Pavlyshenko [2] was published where it was investigated how to construct a regression ensemble of single models using a stacking method. The final output showed that the performance of sales forecasting predictive models can be improved by using the stacking method. In this paper, the stacking method was compared to various models like Extra Tree, ARIMA, Random Forest, Lasso and Neural Network in terms of

Validation Error and Out-of-Sample error where error is minimal in stacking method.

In 2021, Ramachandra et. al. [3] published a paper, where Random Forest Regressor was used. Here various machine learning algorithms are also compared in terms of Mean Absolute Error (MAE) and Mean Square Error (MSE). Also, Accuracy and Root Mean Squared Error (RMSE) are considered for Random Forest Regressor and it achieved an average accuracy of 83.6% and minimum RMSE with value 2829 on the dataset.

In 2021, Dairu and Shilong [4] published a paper where they used XGBoost for sales forecasting. They considered Root Mean Squared Scaled Error (RMSSE) to measure the performance of the model. They have compared XGBoost with Linear Regression and Ridge and found that XGBoost obtained the lower RMSSE of 0.655.

In a paper by Chen et. al. [5], Neural Network Forecasting model was used for Sales prediction. Here Neural Network (NN) forecasting model was compared with Linear Regression and SVM (Support Vector Machine) model in terms of Root Mean Square Error (RMSE) and NN model achieved lowest RMSE (3.20) here.

III. Data Mining Methodology

For the process of the implementation of the forecast, we have opted to follow the CRISP-DM (Cross-Industry Standard Process for Data Mining). We have followed this model for this project since it is the most popular model and provides a structured approach for Data Mining projects. We have covered every phase of the CRISP-DM for the project starting from the Business Understanding to the Evaluation and then to Deployment which is actually the sales prediction.

Business understanding

In the initial phase, we analyzed the datasets available, in order to understand what our problem question would be and how we would be approaching the solution of our question.

Data understanding

In order to understand the dataset even better we visualized the dataset into some of the graphs. Through our visualization, we identified the trends in the dataset in the graph to make the decisions for the choice of the prediction model.

i) We plotted a graph (**Fig 1**) to check the variations of all the features over the years so that we can identify the trends over time. Also observed are the lengths of time for which data are missing especially for the markdown values.

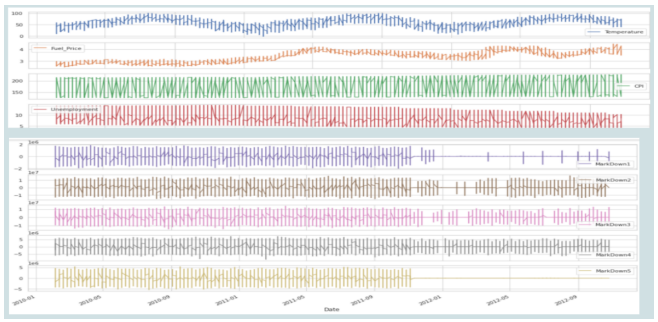

Fig 1: Features vs Time

ii) Then we plotted a bar graph (**Fig 2**) to show the average weekly sales with respect to the store to identify the distribution of sales among all the 45 different stores. We observe that few stores have great sales and few others have an average.

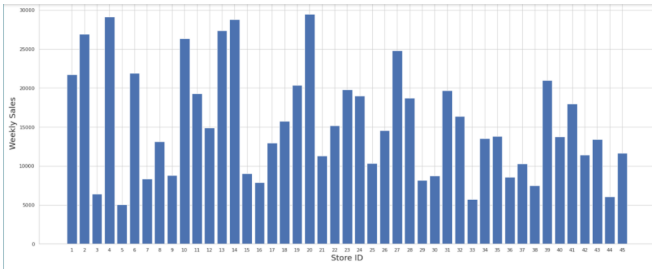

Fig 2: Stores vs Time

iii) Further on, we plotted the graph (**Fig 3**) against the average weekly sales with respect to the departments to understand the distribution of sales among departments. We noticed that a few departments have outperformed in sales. Since we don't have much information about the departments except for the department number, we couldn't explore further wrt departments.

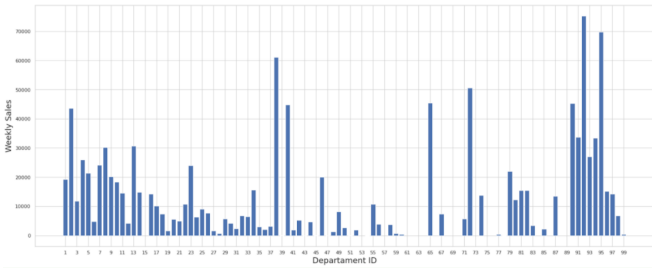

Fig 3: Departments vs Time

iv) We then plotted a boxplot (**Fig 4**) for the weekly sales with respect to the type of stores in order to understand how the sales distribution varies between different types of stores. Lots of outliers are observed in the graph.

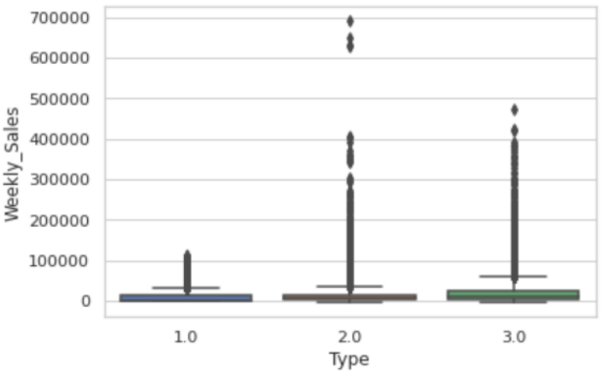

Fig 4: Types vs Time

Fig 5 represents the mean and median values for the weekly sales over the years 2010, 2011 and 2012. One interesting observation to be made is that though the mean is greater than median, both mean and median have the same seasonal component. There is a spike in sales at the end of each year which is the holiday season.

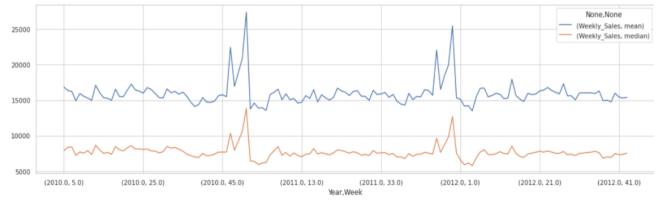

Fig 5: Mean and Median values of Sales

To check the variation of the sales over the duration of 2010 to 2012 we plotted the line graph (**Fig 6**) against the Weekly Sales and the number of weeks. We can see that there isn't a lot of variation and the spike at the end of the year is common for all the years(the holiday spike).

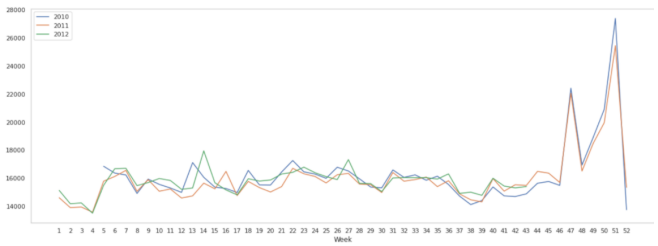

Fig 6: Variation of Sales (2010-2012)

Data preparation

The data pre-processing tasks such as combining the datasets, their transforming, feature-engineering and handling missing values are performed in Python using it's libraries on Google Colab.

Since the IsHoliday and Type columns were in an alphabetical format so we converted them to numerical data. From the data provided by the challenge, it is possible to identify what these holidays are.

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13 --> WEEK 6 Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13 --> WEEK 36 Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13 --> WEEK 47 Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13 --> WEEK 52

We should note that there are no sales records on the labour day holiday at the test dataset since the holiday is in September and the test runs until July.

Null values and Correlations - Exploratory Analysis

The null values we encountered were compared with the total number of values present in the dataset to identify how much of the data is missing. It was observed that MarkDown 1 to 5 has more than 64% of data missing in the dataset. Ideally, since more than 50% of data was missing in these features it is recommended to eliminate them but the MarkDown values identify a certain reduction in sales and imputing values for these features would be the prefered solution.

We then considered the correlations between different features of the dataset to identify how related each feature is to each other using a correlation matrix.

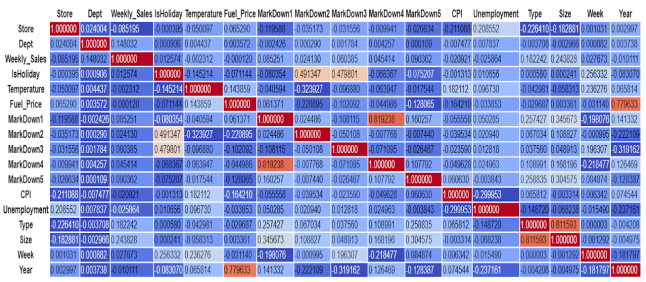

Fig 7: Correlation Matrix of all features

We can observe the following correlations in the above matrix (**Fig 7**)

1. Fuel_Price with Year - Strong correlation
2. MarkDown2 and MarkDown3 with IsHoliday - Moderate correlations
3. MarkDown1 with MarkDown4 - Strong correlation
4. Type with Size - Strong correlation

We can see that Weekly Sales has a weak correlation with other features and it has the highest correlation with Dept, Type and Size.

But we have already established that MarkDown 1 to 5 columns have more than missing values present so we are not looking at the relationships between them.

We shall be performing Multivariate Imputation for the following reasons:

1. A considerable amount of data is missing from multiple variables.
2. Features are not highly correlated and missing data is most probably not MCAR, so univariate imputation was ruled out.

The weekly sales data are grouped by week and year in order to identify the average and median sales per week over the years.The mean values are well above the median, which indicates a high dispersion and variation in sales by stores and departments in a week.

There is a certain pattern over the years, with high seasonality at the end of the year which can be seen in **Fig 5**

The data is grouped by week but separately for each year, in order to identify patterns in weekly sales over the years. As a result, a similar pattern can be seen over the years, with a significant increase in sales in weeks 51 and 47 (Christmas and Thanksgiving). The Superbowl (week 6) and LaborDay holidays (week 36) however have a little impact on increased sales volume.

The data pre-processing tasks such as combining the datasets, their transforming, feature-engineering and handling missing values are performed in Python on Google Colab.

Modelling

We divided our training dataset into a train and test sets by splitting the training dataset in 80:20 ratio respectively.

This is the input for our regression. In order to choose the best model for our prediction, we measure the performance by comparing 8 different training models.

We measure the performance by calculating the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and R Square to check for robustness.

We have analyzed the data using the following models:
Linear Regression - Linear approach to build a relationship between one dependent and one independent variable.
KNN - K-nearest neighbor algorithm is used for regression and classification problems. It uses data and identifies new data based on distance function.
Decision Tree Regressor - In the context of a tree structure, a decision tree constructs regression or classification models. It incrementally breaks down a dataset into smaller and smaller subsets while also developing an associated decision tree. A tree with decision nodes and leaf nodes is the end product.
Random Forest Regressor - A supervised learning algorithm for regression that employs the ensemble learning process. During preparation, a Random Forest constructs multiple decision trees and outputs the mean of the classes as the prediction of all the trees.
Extra Tree Regressor - This class implements a meta estimator that uses averaging to increase predictive precision and control over-fitting by fitting a number of randomized decision trees (also known as extra-trees) on different sub-samples of the dataset. The forest's total number of trees.
XGBoost - XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees.
Ridge - Ridge regression is a model tuning technique that can be used to analyze data with multicollinearity. L2 regularization is achieved using this approach. Where there is a problem with multicollinearity, least-squares are unbiased, and variances are high, the expected values are far from the actual values.
Lasso - A regression analysis approach that includes both variable selection and regularization in order to improve the statistical model's prediction accuracy and interpretability.

IV. Evaluation/Results

The evaluation of the results is done by the calculation of the error rates for each of the models to determine which is the perfect model to implement the forecasting of the "Weekly Sales Prediction".
So we have actually calculated the MAE, MSE, RMSE and R^2 error rates for every model.
The RMSE or the Root Mean Square Error is one of the most widely used metrics to determine the quality of the forecasts. It basically shows how deviated the actual forecasts are from the measured true predictions using the Euclidean Distance. It is seen that the lower the value for the RMSE is going to be the better is the prediction model. [5]

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

A calculation of errors between paired observations describing the same phenomenon is called mean absolute error in statistics. Comparisons of expected versus observed, subsequent time versus initial time, and one measurement technique versus another measurement technique are examples of Y versus X.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

MAE = mean absolute error
y_i = prediction
x_i = true value
n = total number of data points

The sum of the squares of the errors—that is, the average squared difference between the predicted values and the real value—is measured by the mean squared error or mean squared deviation of an estimator in statistics. MSE is a risk function that represents the squared error loss's expected value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

The coefficient of determination, abbreviated R2 or r2 and pronounced "R squared" in mathematics, is the proportion of the variance in the dependent variable that can be predicted by the independent variable.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

The RMSE gives a relatively high weight to large errors since the errors are squared before being averaged. As a result, the RMSE is most useful when significant errors are a concern.

TABLE I. MODEL AND THEIR ERROR

	Model Comparison by Error Rate				
	Name of the Model	***MAE***	***MSE***	***RMSE***	***R^2***
1	Linear Regression	14580.58	479344956.87	21893.9	0.084
2	KNN	14580.17	479344093.44	21893.9	0.084
3	Decision Tree Regressor	2013.40	27530117.92	5246.91	0.947
4	Random Forest Regressor	1492.73	16024665.52	4003.08	0.969
5	**Extra Tree Regressor**	**1498.697**	**16709148.73**	**4087.68**	**0.968**
6	XGBoost	6885.18	138284138.10	11759.42	0.735
7	Ridge	14580.28	479344957.07	21893.94	0.084
8	Lasso	14580.17	479344093.44	21893.9	0.084

CONCLUSION

We have implemented the forecast using the Linear Regression, KNN, Decision Tree Regressor, Random Forest Regressor, Extra Tree Regressor, XGBoost, Ridge & Lasso which have given us different results.

Since we have to decide out of all these models which model is the recommended and the best suited to do the prediction of the Sales Forecast. As per the RMSE values we have concluded the best suited model for the prediction of the “Weekly Sales” is the Extra Tree Regressor model, since it has the lowest error rate of the prediction and high R^2 value, as discussed in [5] and using that model, values have been predicted of the weekly sales accordingly.

REFERENCES

[1] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Southend, UK, 2018, pp. 53-58, doi: 10.1109/iCCECOME.2018.8659115.

[2] B. Pavlyshenko, "Machine-Learning Models for Sales Time Series Forecasting", Data, vol. 4, no. 1, p. 15, 2019. Available: 10.3390/data4010015.

[3] H. V. Ramachandra, G. Balaraju, A. Rajashekar and H. Patil, "Machine Learning Application for Black Friday Sales Prediction Framework," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 57-61, doi: 10.1109/ESCI50559.2021.9396994.

[4] [X. dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021, pp. 480-483, doi: 10.1109/ICCECE51280.2021.9342304.

[5] J. Chen, W. Koju, S. Xu and Z. Liu, "Sales Forecasting Using Deep Neural Network And SHAP techniques," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 135-138, doi: 10.1109/ICBAIE52039.2021.9389930.

[6] Z. Qiao, "Walmart Sale Forecasting Model Based On LightGBM," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2020, pp. 76-79, doi: 10.1109/MLBDBI51377.2020.00020.

[7] M. Singh, B. Ghutla, R. Lilo Jnr, A. F. S. Mohammed and M. A. Rashid, "Walmart's Sales Data Analysis - A Big Data Analytics Perspective," 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Mana Island, Fiji, 2017, pp. 114-119, doi: 10.1109/APWConCSE.2017.00028.

[8] J. Ross Quinlan, "Induction of decision trees", Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[9] Yoav Freund and Robert E Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", computational learning theory, 1995.

[10] J. Ross Quinlan, "C4.5: Programs for Machine Learning" in Morgan Kaufmann, San Francisco, CA, 1993.

[11] C. Cortes and V. Vapnik, "Support vector networks", Machine Learning, vol. 20, pp. 273-297, 1995.

[12] JC. Burges, "A tutorial on support vector machines for pattern recognition. Bell Laboratories", Lucent Technologies, 1997.

[13] JJ, “MAE and RMSE — Which Metric is Better?,” Medium, Mar. 23, 2016. https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d (accessed Apr. 17, 2021).