Store Sales Prediction

Snehaa Durairaj

Abstract:

Background: Accurate store sales prediction is crucial for retailers, impacting inventory management, resource allocation, and overall business success. This project focuses on leveraging machine learning techniques to predict future sales for a retail chain operating in a South American country. By analyzing historical sales data, the project aims to develop a model that can predict future sales.

Methods: The project utilizes a huge dataset spanning almost five years (2013-2017) to train and evaluate the sales prediction model. This project compares machine learning models like Decision Tree, Random Forest, Gradient Boosting, ADA Boosting, and SVM to predict sales, aiming to identify the most accurate model.

Results: Random Forest with feature selection achieved the best performance, with an RMSE of 0.69 and Explained Variance of 0.47. Decision Trees without feature selection also performed well with an RMSE of 0.80. Gradient Boosting had an RMSE of 0.74. However, ADA Boosting and SVM models showed lower performance.

Conclusion: Random Forest Regressor with Wrapper Selection is a promising method of predicting Store Sales. This approach should further be explored with a greater number of features.

1. Introduction

Accurately forecasting store sales remains a constant challenge for retailers. From a consumer's perspective, it is all about stocked and understocked items. While from seller's perspective, it has got lot to with it. The objective of this project is identifying the most accurate sales forecasting method. This, in turn, will empower the retailer to optimize inventory management, make data-driven staffing decisions, and develop targeted promotions, ultimately strengthening their competitive edge.

This project delves into the world of store sales. prediction, through the lens of machine learning for a South American retail chain. I explored four prominent machine learning models – Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machine (SVM) – to analyze historical sales data.

The Initial focus was to ensure the quality of data through a meticulous data cleaning. Once the data was cleansed, I embarked on data preparation tasks to transform the data into a format suitable for machine learning analysis. This included feature engineering techniques to enhance the model's learning ability and interpretability. Moving on with data exploration, both univariant and bivariant

analysis was done to understand variable distributions and relationships. Compelling visualizations communicated insights that were helpful for the results. Finally, machine learning models were built to predict future sales.

2. Literature Review

[1] In a study by Merdas and Mousa (2023), a machine learning model is proposed to predict food sales, emphasizing the influence of feature correlation on prediction accuracy. The model utilizes two datasets, one with high and one with low feature correlation, to train various machine learning algorithms. Ten algorithms are assessed, including Support Vector Machines (SVMs) and Random Forest Regression. Their findings reveal a significant impact of feature correlation on algorithm performance. When dealing with datasets containing high feature correlation, algorithms like SVMs, Lasso Regression, and Bagging Regressor achieved better accuracy. In contrast, datasets with low feature correlation benefitted more from Gradient Boosting, Random Forest Regression, and Decision Tree algorithms.

[2] The study by Andrade and Cunha in 2023 tackles disaggregated demand forecasting, crucial for retailers to manage inventory. Existing research falls short by neglecting real-world complexities. The authors propose a new method using XGBoost, data correction, and structural change detection for improved accuracy. Tested on real data, it outperforms benchmarks and allows for automation. This research offers a solution for complex forecasting and contributes to the field. Future work could explore explanations, alternative models, and better representing marketing effects. By adopting this method, retailers can gain significant advantages.

[3]This article, titled "Sales Forecasting for Retail Chains: A Data Mining Approach", discusses using data mining techniques to predict sales for retail chains. The authors compare different algorithms and find that XGBoost outperforms others. They use a variety of features including store information, customer information, and even weather. This allows them to capture important trends that affect sales. Predicting sales helps stores improve efficiency and revenue.

[4] One of study in 2020, emphasizes the significance of accurate sales forecasting in business and the limitations of traditional methods in handling big data. Data mining techniques are introduced as a solution for improved accuracy and efficiency in sales forecasting.

The authors explore Gradient Boosting Machines (GBM) as a method for sales forecasting due to its ability to handle many features without requiring subjective data reduction. They compare GBM with Random Forest (RF) models and Elastic Net models and find that GBM yields the most accurate results.

[5] This paper from 2018 explores the potential of machine learning for store sales prediction, particularly in the food industry. It argues that machine learning offers advantages over traditional statistical methods due to its ability to utilize powerful algorithms and incorporate external data sources. The review concludes by examining current practices, which often rely on human expertise or basic software solutions using historical averages.

3. Methodology

3.1 3.1 Data Source

A huge dataset comprising of three tables namely, Stores, Sales, and Transactions is being used from the Kaggle website. The "Stores" table provides metadata about each store, such as city, state, type, cluster, and a unique identifier for each store. The "Sales" table records daily sales transactions across various stores with features like Sales, OnPromotion, Date, Family(category of item sold). Meanwhile, the "Transactions" table contains

data on the total number of transactions that took place at each store on a specific date.

3.2 Data Cleaning and Preparation

The initial data cleaning step focused on addressing missing values within the dataset. Fortunately, the dataset contained no duplicate entries. Since the dataset was large, and the number of missing values was relatively small (only 20 rows), I decided to remove these rows with missing values since the impact on the overall data analysis was considered minimal.
To simplify analysis, three separate tables (Stores, Sales, Transactions) were merged into a single dataframe. This unified view combined information about stores, daily sales, and transactions, gave a more detailed exploration of sales patterns.

3.3 Feature Engineering

Firstly, the "Family" attribute within the sales data was addressed, which presented a challenge due to the high number of categories (around 33). Many of these categories exhibited a high degree of similarity, such as "Grocery I" and "Grocery II." To address this redundancy and improve model efficiency, I merged highly correlated categories into a more general category. Through this process, I successfully reduced the total number of categories to just 10. Furthermore, the original sales data stored dates in a single format (dd/mm/yyyy). To gain a deeper understanding of sales trends across various timeframes, this single date column was split into three distinct columns: day, month, and year.
In addition, I performed [7]label encoding on all categorical variables to convert them into numerical values for analysis purposes. These encoded variables were stored separately without altering the original dataset, ensuring that the visualization process remained unaffected.

3.4 Feature Selection:

[10]During the feature selection stage, I identified eleven variables as input for the initial analysis of the models. However, the dataset presented a challenge, containing only eleven features after excluding irrelevant ones like 'store_nbr' and 'id'. Due to this limitation, applying a traditional feature selection technique wasn't feasible. Consequently, all eleven features were used as inputs.

3.5 Data Visualization:

To gain a deeper understanding of the sales data and identify key factors influencing sales, the data visualization approach was used. This process involved both univariant and bivariant analysis.

Univariate Analysis: Visualizations like histograms and box plots to examine the distribution of individual features. This helped us to identify couple of outliners in the data.

Bivariant Analysis: Scatter plots and heatmaps were constructed to explore the relationships between pairs of variables.

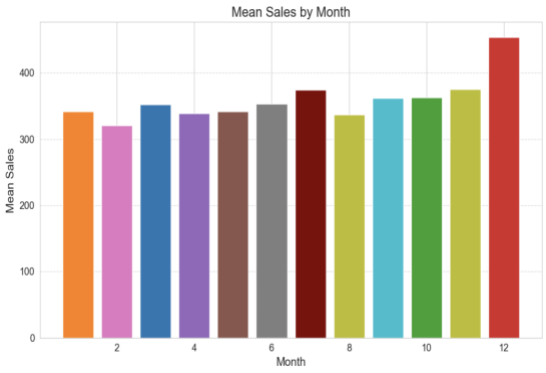

Fig 3.5.1: Mean Sales by Month

From the figure 3.4.1, sales data shows a trend of higher sales in December and July compared to other months. This pattern coincides with vacation periods. December, being a festive time, and July, coinciding with summer and

good weather, might encourage increased consumer spending and leading to higher sales.

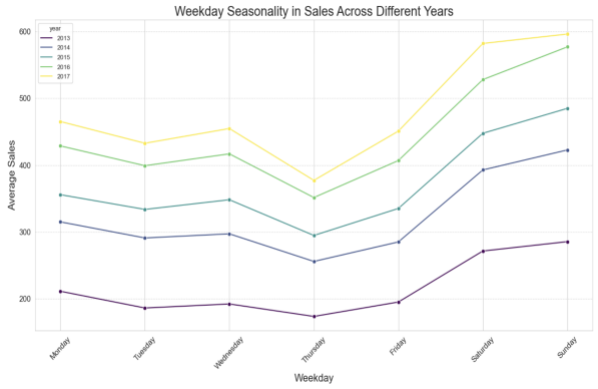

Fig 3.5.2 Weekday Seasonality in Sales across Different Years

The Fig 3.5.2 clearly shows that maximum sales happen during the weekends and suggests a positive trend in the average daily sales across the days, with 2017 exhibiting the highest value. This indicates an increase in consumer spending year-over-year. However, factors like price inflations could also play an important role in this trend.

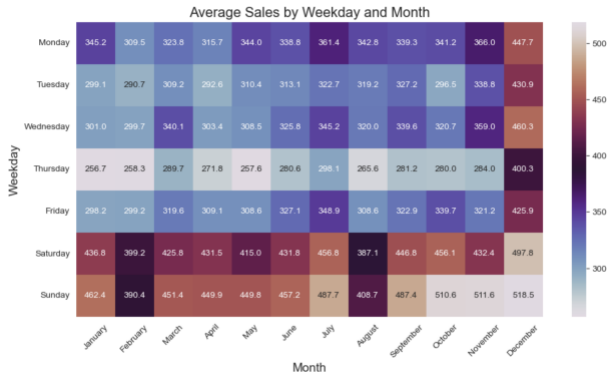

Fig 3.5.3 Average Sales by Weekday and Month

Fig 3.5.3 reveals that, analysis of average sales by weekday across months reveals a distinct pattern. Weekends, particularly Saturdays and Sundays, see the highest sales. Mondays also show a notable increase in sales compared to other weekdays. Additionally, December appears as the month with the highest overall sales.

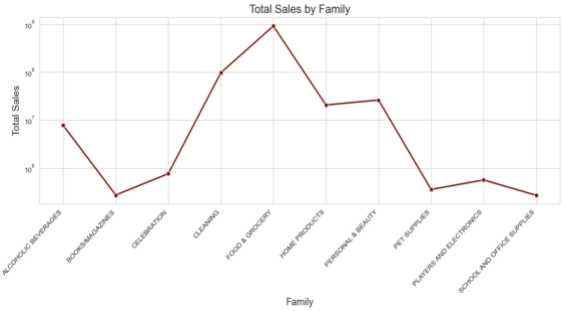

Fig 3.5.4 Total Sales By Family

Fig 3.5.4 figure reveals that the "Food & Groceries" category generates the highest sales among all family categories, followed by the "Cleaning" category. Conversely, "Books & Magazines" and "School & Office Supplies" see the lowest sales.

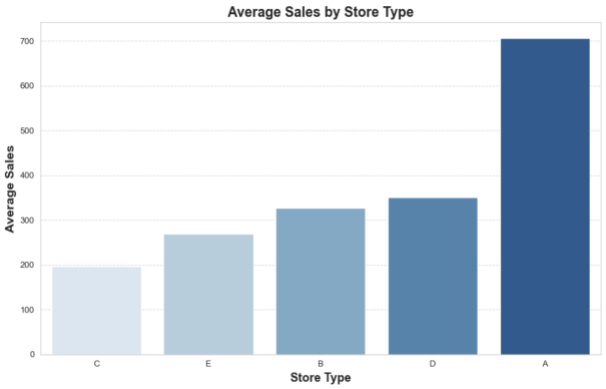

Fig 3.5.5 Avg Sales by Store Type

The Fig 3.5.5 shows a trend where Store Type A generates higher sales compared to other store types. Based on this finding, retailers might consider to strategically increase the number of Type A stores in the future.

3.6 Model Building

For model building in this project, I chose a combination of non-parametric and parametric regression techniques. Non-parametric models, such as Decision Trees, Random Forests, Gradient Boosting, and AdaBoost, were used

due to their flexibility in handling various data distributions. Additionally, Support Vector Regression (SVR) model, which is a parametric approach, was also included in the analysis.

3.6.1 Decision Tree Regressor

[8]Decision Tree Regression is a machine learning algorithm adept at predicting continuous numerical values. The algorithm begins by analyzing the entire dataset and identifying the most critical feature (variable) that can effectively split the data into two distinct groups. This process continues recursively, creating a branching structure resembling a tree. At each junction, the tree poses a question about a specific feature, further dividing the data based on the answer. This iterative questioning continues until reaching "leaves" at the end of the branches. These leaves represent specific scenarios and finally leading to result.

3.6.2 Random Forest Regressor

[1] [4] [9]Random Forest Regressor is a supervised machine learning algorithm used for predicting continuous numerical values. This operates by creating a "forest" of numerous decision trees, each acting as an independent predictor. The trees in random forests run in parallel, meaning there is no interaction between these trees while building the trees. The final prediction is an average of the individual tree predictions.

3.6.3 Gradient Boosting Method

[2][3][4]Gradient boosting is a powerful machine learning technique that works by combining multiple simple predictive models sequentially, with each new model correcting errors made by the previous ones. This iterative approach focuses on minimizing prediction errors, employing a gradient descent optimization algorithm to optimize the overall predictive performance. Gradient boosting has gained popularity due to its ability to handle complex data patterns and produce highly accurate predictions including both regression and classification tasks.

3.6.4 ADA Boosting Method

[6]AdaBoost is an ensemble method that sequentially trains and deploys trees to form a robust learner. By implementing boosting, a series of weak classifiers is connected, with each aiming to refine the classification of samples misclassified by its predecessor. This iterative approach fosters adaptability, concentrating on challenging instances to enhance overall classification accuracy.

3.6.5 Support Vector Regression (SVR)

[1] Support Vector Regression (SVR) is a machine learning algorithm used for regression tasks. It is based on the principles of Support Vector Machines (SVM), which are widely used for classification. SVR works by finding a hyperplane in a high-dimensional space that best represents the relationship between the input features and the target variable. SVR aims to minimize the error while still staying within a specified margin. This also handles non-linear relationships and outliers using the kernel functions.

Models	RMSE	Explained Variance
Decision Tree	0.80	0.36
Decision Tree with Wrapper Select	0.86	0.17
Random Forest	0.70	0.46
Random Forest with Wrapper Select	0.69	0.47
Gradient Boosting	0.74	0.41
ADA Boosting	0.86	0.18
SVM with 'rbf' kernel	0.88	0.15
SVM with 'linear kernel	0.90	0.10
SVM with 'sigmoid' kernel	0.91	0.09

Table 4.1: Sales Prediction Model Performance

4. Results:

I assessed the models based on two key metrics: Root Mean Squared Error (RMSE) and Explained Variance (Expl Var). Lower RMSE indicates better model performance, signifying a closer fit between predicted and actual sales values. Conversely, higher Explained Variance means, a greater proportion of the variance in the sales data was explained by the model.
The four models were run against different fine-tuning parameters, I have listed down the parameters that gave me the best results.

Random Forest with feature Selection achieved the overall best performance with RMSE 0.69 and Expl. Var of 0.47. This model's parameters were: {criterion='mse', splitter='best', max_depth=None, min_samples_split=3, min_samples_leaf=1, max_features=None, random_state=1}

The Random Forest model, without feature selection gave the next lowest RMSE (around 0.7) and an Explained Variance of approximately 0.46 with parameters: {n_estimators=100, max_features=0.33, max_depth=None, min_samples_split=3, criterion='mse', random_state=1}

The Decision Trees without feature selection gave better results of RMSE 0.80 and explained of 0.36 when compared to with feature selection(RMSE: 0.86 and Expl. Var 0.17) the parameters for Decision tree are: {criterion='mse', splitter='best', max_depth=None, min_samples_split=3, min_samples_leaf=1, max_features=None, random_state=1}

Decision Tree and Random Forest with Feature Selection selected the same the subset of three features (namely "On Promotion," "Family," and "Day") resulted in very minimal changes in the RMSE and Explained Variance.

The Gradient Boosting model achieved an RMSE of 0.74 and Explained Variance of 0.41 with parameters : n_estimators=100, loss='ls', learning_rate=0.1, max_depth=3, min_samples_split=3, random_state=1).fit(data_train, target_train, demonstrating its ability to learn from the data and make reasonably accurate predictions.

ADA Boosting and SVM models yielded lower performance. Compared to the above models, ADA Boosting and SVM models with different kernels (rbf, linear, sigmoid) and {gamma=0.1 and C=1.0} did not perform well and resulted in higher RMSE (above 0.85) and lower Explained Variance (below 0.18). This suggests these models might not have captured the underlying patterns in the sales data as effectively as others.

5. Discussion

This study explored the application of various machine learning models for store sales prediction. By evaluating model performance using Root Mean Squared Error (RMSE) and Explained Variance, I aimed to identify the most effective approach for capturing the complex relationships between influencing factors and sales.

The analysis revealed Random Forest, particularly with feature selection, as the leading model with the lowest RMSE (0.69) and a significant explained variance of 47%. This suggests that Random Forest effectively learned the intricate relationships between various features and sales, enabling accurate predictions. Interestingly, even without feature selection, Random Forest delivered a good performance, highlighting its robustness and potential.

Overall, this research successfully demonstrated the feasibility of machine learning techniques for store sales prediction. By continuing to explore advanced methods, I'm planning incorporate richer datasets, and delve deeper into model interpretability. Future research can pave the way for even more robust and insightful sales prediction models, ultimately empowering retailers to optimize their operations and maximize profitability.

6. Conclusion:

In this study, I have used four Machine Learning techniques namely, Decision Tree, Random Forest, Gradient Boosting, Ada Boosting, and SVR to predict sales, aiming to identify the most accurate model.

Just by understanding how sales have changed over time, businesses gain valuable insights. This knowledge empowers them to make smart decisions that can boost sales and keep them competitive in the marketplace. They can use this information to identify these trends, predict future sales patterns, and ultimately develop plans for their financial management that ensure their success.

7. Future Work

This initial analysis has laid a foundation for predicting store sales. However there are couple of items I'm planning to incorporate in the future.

The current dataset utilizes a limited number of features. Incorporating additional relevant features, such as weather data (e.g., Sunny, Snowy, Winter, Rain), holidays, and promotional activities, could potentially improve the model's ability to capture the factors influencing sales.

As the dataset grows, leveraging big data tools like Apache Spark for distributed computing could become necessary to efficiently manage the data processing and model training pipeline.

The proposed revisions would create a more structured and comprehensive framework for sales prediction, ultimately empowering retailers with the ability to forecast sales with greater ease and accuracy.

8. References

[1] Merdas, Hussam & Hameed Mousa, Ayad. (2023). Food sales prediction model using machine learning techniques. International Journal of Electrical and Computer Engineering (IJECE). 13. 6578. 10.11591/ijece.v13i6.pp6578-6585.

[2] Luiz Augusto C.G. Andrade, Claudio B. Cunha,
Disaggregated retail forecasting: A gradient boosting approach,
Applied Soft Computing,
Volume 141,
2023,110283, ISSN 1568-4946,
https://doi.org/10.1016/j.asoc.2023.110283.

[3] Jain, A., Menon, M. N., & Chandra, S. (2015). Sales forecasting for retail chains. *San Diego, California: UC San Diego Jacobs School of Engineering*.

[4] Antipov, E.A., Pokryshevskaya, E.B. Interpretable machine learning for demand modeling with high-dimensional data using Gradient Boosting Machines and Shapley values. *J Revenue Pricing Manag* 19, 355–364 (2020). https://doi.org/10.1057/s41272-020-00236-4

[5] Tsoumakas, G. A survey of machine learning techniques for food sales prediction. *Artif Intell Rev* **52**, 441–447 (2019). https://doi.org/10.1007/s10462-018-9637-z

[6] Peng Yang, Shiguang Shan, W. Gao, S. Z. Li and Dong Zhang, "Face recognition using Ada-Boosted Gabor features," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings., Seoul, Korea (South), 2004, pp. 356-361, doi: 10.1109/AFGR.2004.1301556.

[7] A. Mottini and R. Acuna-Agost, "Relative Label Encoding for the Prediction of Airline Passenger Nationality," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016, pp. 671-676, doi: 10.1109/ICDMW.2016.0100.

[8] Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20 - 28. https://doi.org/10.38094/jastt20165

[9] Yile Ao, Hongqi Li, Liping Zhu, Sikandar Ali, Zhongguo Yang,
The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling,
Journal of Petroleum Science and Engineering,
Volume 174,
2019,
Pages 776-789,
ISSN 0920-4105,
https://doi.org/10.1016/j.petrol.2018.11.067.

[10] A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.