Sales Prediction in Ecuadorian Stores

By Shuhan Li and Haoran Wu

Background

The retail industry operates in a highly competitive environment where success is significantly influenced by the ability to predict future sales accurately. Sales forecasting not only aids in optimizing inventory management but also assists in making informed decisions about staffing, marketing, and financial planning. The advent of machine learning and data analytics has provided retailers with powerful tools to enhance their forecasting accuracy, thereby leading to improved operational efficiency and increased profitability.

In recent years, the accumulation of extensive historical sales data, alongside advancements in computational technologies, has enabled more sophisticated analytical approaches. Retail giants have increasingly relied on predictive analytics to foresee consumer demand patterns and adjust their strategies accordingly. This project taps into these advancements by utilizing a comprehensive dataset from Kaggle's "Store Sales - Time Series Forecasting" competition. This dataset includes crucial variables such as store information, product details, and historical sales data across several store locations and products.

Problem Description

The primary challenge addressed in this project is the accurate prediction of store sales using historical data. Predicting sales is inherently complex due to the multitude of factors that can influence outcomes, such as economic conditions, store promotions, competition, local events, and seasonal trends. Furthermore, the performance of sales forecasting models can significantly impact a retailer's bottom line—underestimating sales can result in stock shortages and lost sales, while overestimating them can lead to excess inventory and increased costs.

The dataset provided includes daily sales data, which offers a granular look at sales trends and patterns but also introduces challenges due to its voluminous size and the noise inherent in day-to-day operations. Additionally, the dataset encompasses multiple products and store locations, each with potentially unique characteristics and sales drivers that need to be accounted for in the predictive modeling process.

The objective of this project is therefore to develop a robust model that can effectively utilize the available data to forecast sales accurately. This involves handling a large volume of data, selecting appropriate features, choosing the right modeling techniques, and validating the models to ensure they are reliable and applicable in a real-world retail setting.

Analysis

Dataset Description:

The dataset for this project comes from the "Store Sales - Time Series Forecasting" competition hosted on Kaggle, focusing on the retail sales data of Favorita, a chain of stores located in Ecuador. The goal of this analysis is to predict future sales across various product families within these stores using historical data enriched with additional contextual information.

The primary training data comprises time series records, capturing daily sales figures, store details, product categories, and promotional activity. Each entry details the sales for a specific product family at a given store on a particular date, including whether any items were on promotion that day. In addition to the core sales data, the dataset includes several auxiliary files that provide a richer context for the sales data:

Store Information file includes its location (city and state), type, and cluster identifier that groups stores with similar characteristics. The Economic Indicators provided daily oil prices reflecting the economic backdrop of Ecuador, a country whose economy is sensitive to oil price fluctuations. This data is crucial as shifts in economic health can significantly impact consumer spending patterns. The last Calendar Events present the details of holidays and significant events, which are essential for accounting for spikes or drops in sales. This file also notes any holidays that have been moved or extended, affecting the usual patterns of consumer behavior during these periods.

Data Preprocess:

To enhance our predictive model with additional features, we undertook a comprehensive data integration process. We combined the main training dataset with several supplementary datasets to enrich the predictors available for our analysis. This included merging daily oil prices as an economic indicator and holiday events to mark special days, both aligned by date. Additionally, we integrated store-specific information using the store identification numbers.

For more efficient processing and analysis within our models, we converted the date fields from string formats to pandas.DateTime objects. This not only facilitates time-series analysis but also helps in handling temporal data more effectively.

Moreover, we paid particular attention to holidays that were officially transferred to different dates by the government. To accurately reflect their impact on sales, we calculated the actual celebration dates of these transferred holidays. Typically, a transferred holiday is treated more like a standard day, except for its new designated date, which carries the significance of the holiday. This nuanced handling of dates ensures our model accurately reflects real-world shopping behavior around holidays and special events.

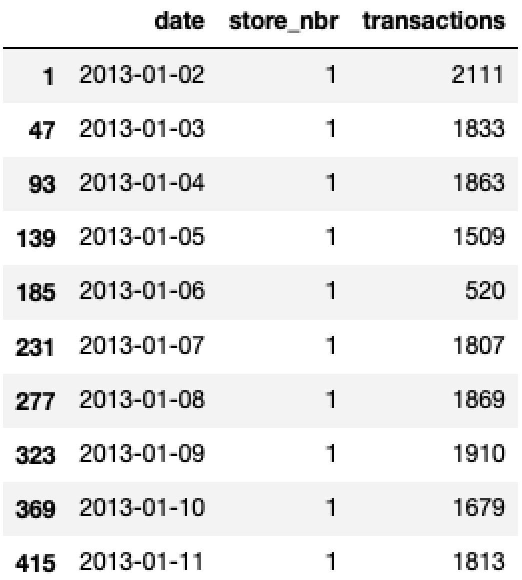
For our CNN-LSTM model, a critical component of the data preprocessing involved transforming categorical features such as 'city', 'store type', and 'cluster' into formats suitable for neural network processing. To achieve this, we implemented a data pipeline that first applies a continuous target encoder to these categorical variables. This method transforms categorical entries into numerical values based on the mean target value (sales) associated with each category. This encoding helps the model capture the impact of these categories on sales predictions more effectively.

Following the target encoding, we utilized the KBinsDiscretizer to categorize these numerical representations into discrete bins. This transformation groups the encoded features into bins, turning them into unique categorical feature groups. This step is particularly useful for handling features with a wide range of values by clustering similar values together, which simplifies the model’s learning process.

For the CNN-LSTM model, to convert category features, like 'city', 'typestores', and 'cluster', we created a data pipeline using the continuous target encoder and category them into bins using KBinsDiscretizer to identify each category as a unique feature group. We also checked missing values and filled them by using Interpolation.

Exploratory Data Analysis:
There are 54 stores and 33 prodcut families in the data. The time serie starts from 2013-01-01 and finishes in 2017-08-31. The dates in the test data are for the 15 days after the last date in the training data. Date range in the test data will be very important to us while we are defining a cross-validation strategy and creating new features.

Transactions:



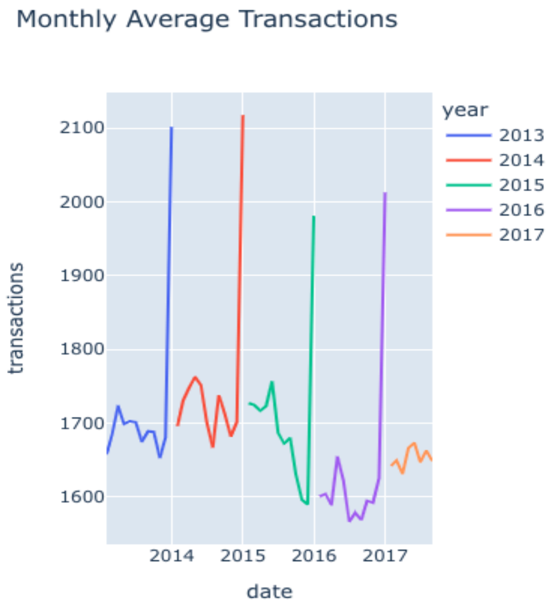
	date	store_nbr	transactions
1	2013-01-02	1	2111
47	2013-01-03	1	1833
93	2013-01-04	1	1863
139	2013-01-05	1	1509
185	2013-01-06	1	520
231	2013-01-07	1	1807
277	2013-01-08	1	1869
323	2013-01-09	1	1910
369	2013-01-10	1	1679
415	2013-01-11	1	1813

This feature is highly correlated with sales but first, you are supposed to sum the sales feature to find relationship. Transactions means how many people came to the store or how many invoices created in a day.

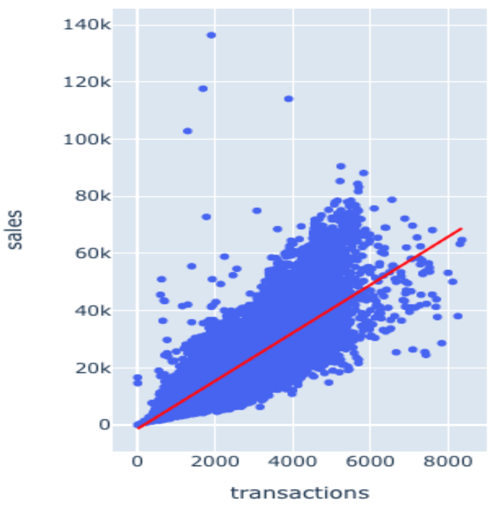
Sales gives the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).

That's why, transactions will be one of the relevant features in the model.

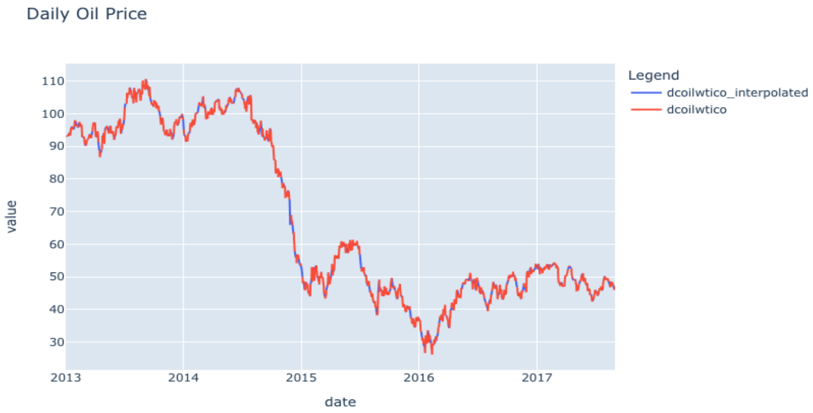
Let's take a look at transactions by using monthly average sales!



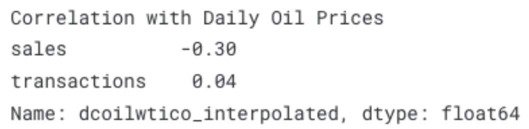
When we look at their relationship, we can see that there is a highly correlation between total sales and transactions also.



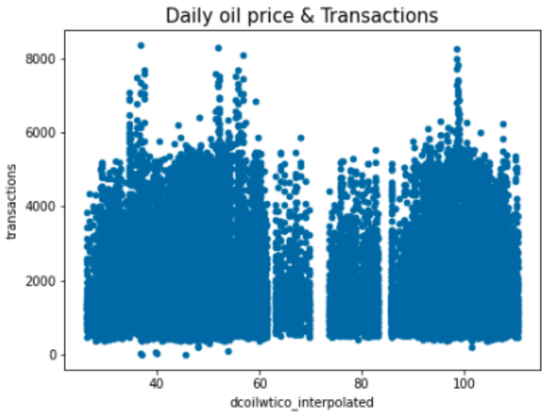
Oil:
In our case, Ecuador is an oil-dependent country. Changing oil prices in Ecuador will cause a variance in the model.

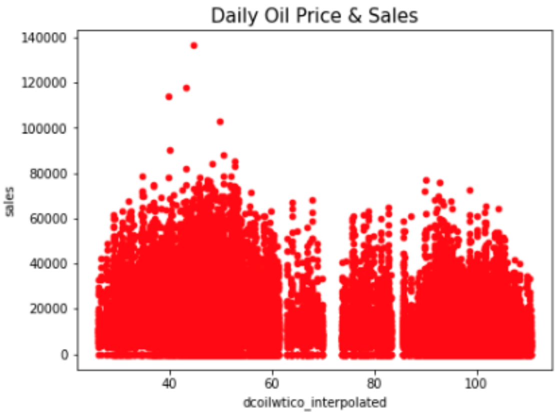


First of all, let's look at the correlations for sales and transactions. The correlation values are not strong but the sign of sales is negative. Maybe, we can catch a clue. Logically, if daily oil price is high, we expect that the Ecuador's economy is bad and it means the price of product increases and sales decreases. There is a negative relationship here.



```
Correlation with Daily Oil Prices
sales          -0.30
transactions    0.04
Name: dcoilwtico_interpolated, dtype: float64
```

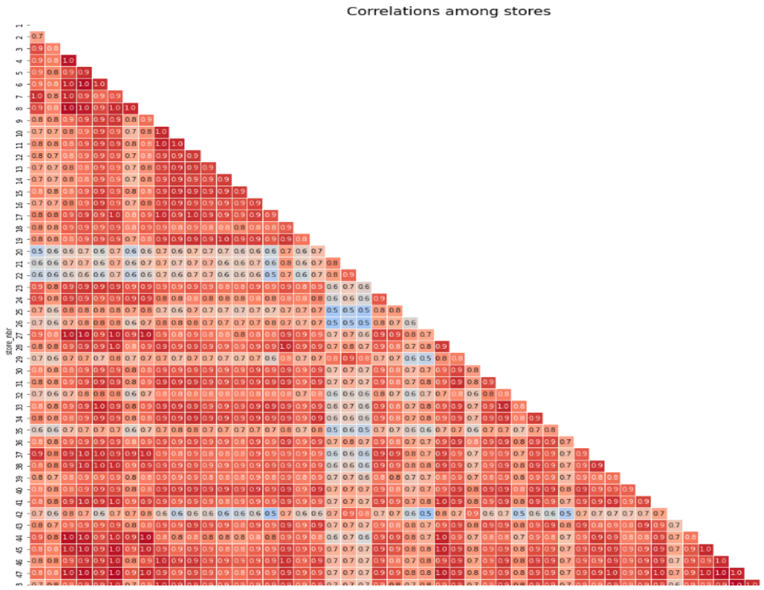


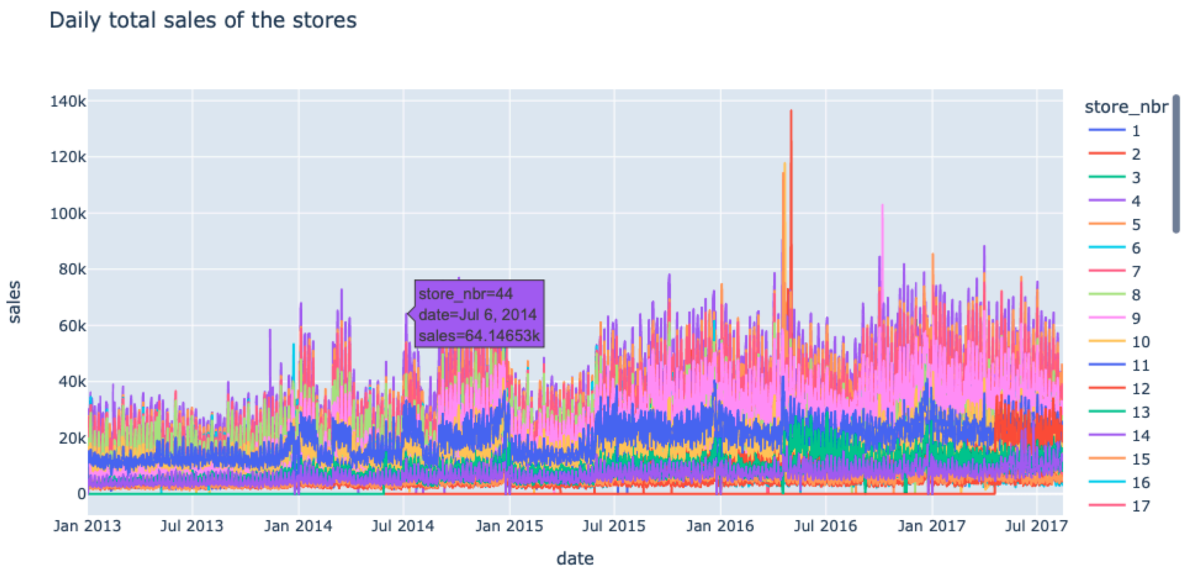


Sales:

Our main objective is, predicting store sales for each product family. For this reason, sales column should be examined more seriously. We need to learn everthing such as seasonality, trends, anomalies, similarities with other time series and so on.

Most of the stores are similar to each other, when we examine them with correlation matrix. Some stores, such as 20, 21, 22, and 52 may be a little different.





In this project, we are going to apply two different modes to make sales predictions.

Linear Regression:
We just use the linear regression model as a simple model to do the basic predictions. Obviously it's not really good for this project.

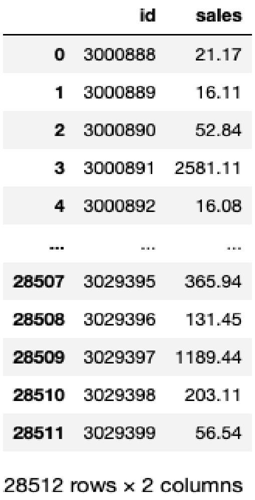
CNN-LSTM:

For this project, we created several models for each feature category bin and employed a hybrid neural network model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) to forecast sales. The convolutional layer serves as the feature extractor. It processes input data through convolutional filters, which helps in capturing the spatial relationships within the data, such as patterns across different store attributes and product families. This is crucial for understanding complex interactions that can affect sales, like simultaneous promotions of multiple products. The output from the convolutional layer is passed to an LSTM layer. LSTMs are designed to recognize patterns in sequences of data and are particularly effective for predictions where past information, like previous days' sales and events, is crucial for forecasting future events. The LSTM layer analyzes these temporal features to predict sales trends over time.

To optimize the performance of our CNN-LSTM model, we committed to an extensive training regimen, running the model through 100 epochs. This extended training period was crucial to ensure the model thoroughly learned from the complexities and nuances of the dataset. However, to guard against overfitting we incorporated regularizers to penalize overly complex models, implemented dropout layers to randomly deactivate a fraction of neurons during training which helps in generalizing the model better, and utilized early stopping to halt training when the validation performance ceased to improve. These techniques collectively help maintain the model's generalizability to new, unseen data while allowing it to learn deeply from the training data.

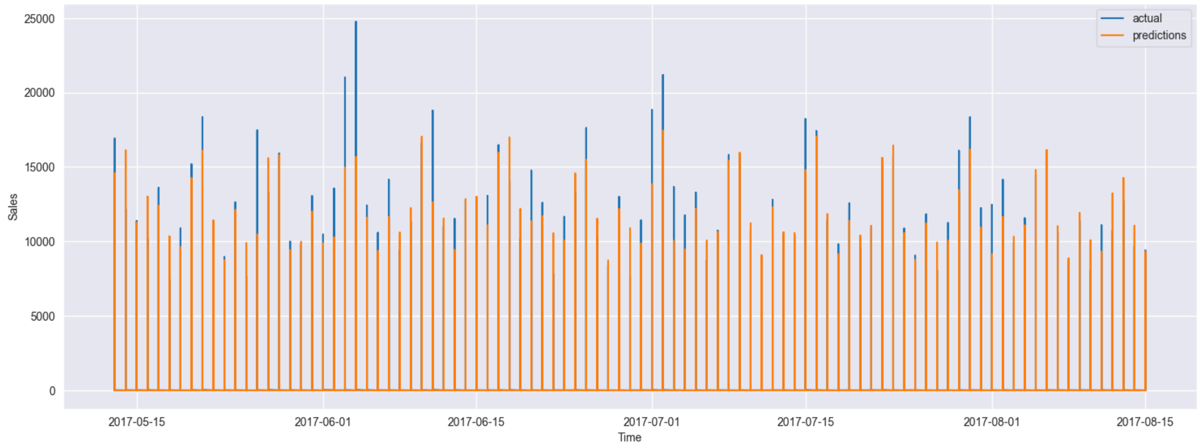
Results

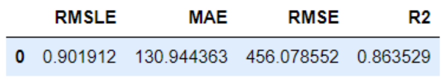
Linear regression:



	id	sales
0	3000888	21.17
1	3000889	16.11
2	3000890	52.84
3	3000891	2581.11
4	3000892	16.08
...	...	...
28507	3029395	365.94
28508	3029396	131.45
28509	3029397	1189.44
28510	3029398	203.11
28511	3029399	56.54

28512 rows × 2 columns





	RMSLE	MAE	RMSE	R2
0	0.901912	130.944363	456.078552	0.863529

The graphical representation of actual versus predicted sales over the time frame from mid-May to mid-August 2017 displays a close alignment between the two datasets, showcasing the model's proficiency in tracking sales trends.

While the actual sales exhibit periodic spikes—which likely correspond to promotional events or holidays—the predicted sales generally follow the same pattern, capturing the essence of these fluctuations. These variances are particularly pronounced in instances where promotional activities or holidays lead to significant deviations from typical sales patterns.

The Root Mean Squared Logarithmic Error(RMSLE) of 0.901912 suggests that the model's predictions are in reasonable proportion to the true sales values, especially given the logarithmic nature of this metric which penalizes underestimates more than overestimates. The Mean Absolute Error(MAE) stands at 130.944363, offering insight into the average magnitude of the errors in the predictions. The Root Mean Squared Error (RMSE) is observed at 456.078552, providing a measure of the typical deviation from the actual sales figures. The R^2 value of 0.863529 is indicative of a high level of variance being accounted for by the model.

These results collectively affirm the effectiveness of the CNN-LSTM model in grasping the complex patterns within the sales data. While there is room for improvement, particularly in better capturing peak sales days, the overall performance demonstrates a robust foundation for making informed business decisions based on the model's sales predictions. Further tuning and additional feature engineering could potentially refine the model's predictions, further closing the gap between predicted and actual sales.

Conclusion

Our prediction project demonstrates that the CNN-LSTM model effectively captures the complex dynamics of retail sales data, delivering robust forecasts that aid in strategic decision-making. The model's proficiency is evident from its alignment with actual sales trends, effectively capturing fluctuations due to promotional events and holidays. Despite occasional deviations during peak sales days, the performance metrics such as RMSLE, MAE, RMSE, and R^2 indicate a high degree of accuracy in the forecasts, underscoring the model’s capability to navigate the intricacies of retail sales forecasting.

The project also underscores the importance of meticulous data preprocessing and feature engineering, which significantly enhance model performance. In preparation for training, we employed a robust data pipeline along with categorical encoding techniques to ensure the data was optimally formatted and ready for model ingestion. To safeguard against overfitting, we applied a combination of strategies, including the integration of early stopping mechanisms, the utilization of regularization methods, and the incorporation of dropout layers within the network architecture. These measures are designed to enhance the model's ability to generalize and perform reliably on unseen data. By integrating diverse datasets such as economic indicators and holiday schedules, the model can more accurately reflect the factors influencing consumer behavior. The effective use of CNNs to extract spatial relationships and LSTMs to analyze temporal patterns exemplifies the potential of hybrid approaches in predictive analytics.

Moving forward, there is potential for further refinement of the model by expanding feature sets and tuning model parameters to even better capture peak sales events. The success of this project lays a solid foundation for future exploration and optimization in the field of sales forecasting, providing a valuable tool for retailers seeking to optimize operations and increase profitability.