IST 707 - APPLIED MACHINE LEARNING

"FINAL PROJECT REPORT"

TITLE: Inventory Demand Forecasting

TEAM:

- Tejas Mistry (GitHub ID: tamistry)
- Deep Patoliya (GitHub ID: deep2312)
- Kunal Jain (GitHub ID: kunal7387)
- Kruti Kotadia (GitHub ID: krutikotadi)

INRODUCTION

This project has addressed a limitation in the ability of grocery chains to efficiently optimize inventory management to fluctuations in demand due to changing weather patterns caused by climate change. Our primary stakeholders are the inventory managers of these chains, who need to optimize sales and avoid surpluses and avoid selling out of items. By hosting historical sales data against weather conditions, we have created predictive models using absolute values ARIMA and relative values i.e., difference from the previous day and the same day the previous week LSTM to predict sales fluctuations caused by weather. Our results have concluded a clear correlation between weather and sales in this data, allowing inventory managers to optimize inventory management by adjusting the levels of various stock of data-driven decisions. This reduces waste and maximizes profitability.

LITERATURE REVIEW

Research has shown that weather significantly impacts consumer purchasing behaviors, especially in grocery retail, where product sales like beverages and comfort foods fluctuate with weather changes (Murray and Di Muro, 2010). To address inventory management challenges for grocery chains, we employed ARIMA and LSTM models, known for their robust forecasting capabilities. ARIMA excels at modeling non-stationary and seasonal data, suitable for linear sales trends related to weather (Box and Jenkins, 1976), while LSTM networks manage long-term dependencies, capturing how past weather conditions influence future sales (Hochreiter & Schmidhuber, 1997). These models have proven effective across various sectors, supporting our approach to improve inventory efficiency and profitability in grocery chains (Ediger & Akar, 2007; Alon et al., 2001).

DATA AND METHODS

The project utilizes a detailed dataset which contains two primary elements: detailed daily sales data from a regional grocery store chain as well as corresponding weather information that was sourced from a trusted national meteorological organization. By combining these two sets of information, we are able to investigate how buying decisions within the food market are affected by climatic conditions.

Sales Data:

The Sales data contains over 100,000 records, spanning from January 2022 to March 2024. Each record in this dataset provides comprehensive transactional details including the date of the transaction, product descriptions, quantities sold, and sale prices. It encompasses a diverse range of product categories, thereby offering a granular view of consumer purchasing patterns under various weather conditions. The dataset is structured with numerous attributes:

- Date: The date on which the transaction occurred.
- Product Description: A textual description of the product sold.
- Quantity Sold: The number of units of the product sold.
- Sales Price: The total revenue generated from the transaction.
-

Weather Data

The weather dataset aligns with the sales data timeframe and includes daily meteorological measurements such as average, maximum, and minimum temperatures, precipitation levels, wind speed, and other relevant weather variables. This dataset is integral for our analysis as it allows us to correlate specific weather patterns with fluctuations in grocery sales. The high reliability of this data is guaranteed by the stringent data collection and validation standards upheld by the national meteorological service.

Data Quality and Challenges

Data preparation was a significant challenge because our sales data was not evenly distributed among the various product categories. This issue was exacerbated by the seasonal item category having extreme variations and may lead to biases when analyzing it. This led us into applying normalization methods as a way of balancing the dataset. We also had few days when we did not have any weather condition records available in our database; hence there were gaps that we filled using interpolation techniques so that we could have continuity.

Exploratory Data Analysis (EDA)

Our exploratory data analysis was comprehensive, involving various statistical and graphical techniques to understand the underlying patterns and relationships within the data. We created a range of visualizations to aid in this analysis:

- Histograms and Box Plots: Used to examine the distribution and variance of sales across different weather conditions.
- Scatter Plots: Developed to scrutinize the correlations between weather variables like temperature and specific product sales, revealing trends such as increased ice cream sales during higher temperature days.
- Time Series Plots: These plots were particularly useful in identifying seasonal trends and the effect of specific weather phenomena on sales activities.

We analyzed how weather conditions affect grocery sales by using data preprocessing methods and advanced predictive analytics approaches. We selected methods that could predict sales accurately from time series data because such data contain both linear and non-linear patterns. These patterns are due to a variety of factors, such as weather.

Data Cleaning and Imputation: Our initial step involved rigorous cleaning of the datasets to ensure accuracy and consistency. In the sales data, missing entries for product quantities and sales were imputed based on median sales figures of similar product types during comparable periods, which preserved the integrity of seasonal patterns. Weather data, often missing critical metrics like temperature or precipitation, was interpolated using a time-series-specific method that considers the temporal proximity of available data points to maintain natural weather progression.

Feature Engineering: To enrich our models, we engineered several features that reflect potential influences on sales patterns:

- Temporal Features: We added time-related features such as day of the week, month, and special holiday flags to capture periodic sales fluctuations.
- Weather-derived Features: Features like rolling averages of temperature and cumulative precipitation over the week preceding each sale day were created to account for the delayed effects of weather conditions on consumer behavior.

Normalization and Transformation: Numerical features were normalized using standard scaling techniques to facilitate faster convergence during model training. Categorical variables, particularly product descriptions, were transformed using one-hot encoding to allow for integration into our predictive models.

Predictive Modeling

ARIMA Model Implementation

The Autoregressive Combined Moving Average (ARIMA) Representation is used because it is extremely good at Forecasting Information that shows seasonal Layouts or trends which are important aspects of our weather-influenced sales Information. The Forecast package's auto. Arima() Role performs a grid search to minimize the Akaike Information Criterion (AIC) determining the optimal values for the moving average term (q) degree of differencing (d) and autoregressive term (p) of the ARIMA Representation.
This technique ensures that the representation accurately captures seasonality and underlying non - stationarity in information offering a good fit to help forecast sales patterns for layouts that depend on outdoor variables such as weather.

LSTM Neural Networks

Alternatively, Long Short-Term Memory (LSTM) Webs are employed to manage the Complicated non-linear interactions between weather conditions and sales Information over extended periods. Our LSTM Structure includes multiple layers with dropout regularization to mitigate overfitting and it is Improved using the Adam Improver which adjusts learning rates based on individual Effectiveness Improvements enhancing Teaching Productivity.
Developing an LSTM Representative strategy that models sales data daily but also takes into account weather data entails learning from long-term dependencies that exist within the Information set. By combining ARIMA, which captures linear relationships as well as seasonal variation, with LSTM's ability to model broader and more complex patterns, we can create a well-rounded forecast model. This two-pronged method not only enhances the accuracy of our sales predictions but also provides deep insights into them.

Methodological Challenges and Adjustments

Our initial efforts in representation using simpler techniques such as moving averages or exponential smoothing were not sufficient as they have less ability to handle complex and multi-faceted data sets. For example, these representations did not capture how various kinds of products are affected concurrently by weather conditions in several ways.

After evaluating these initial Representations, we pivoted to a combination of ARIMA and LSTM Representations. ARIMA was specifically useful for short-term forecasting and adjusting inventories on a weekly basis while LSTM provided valuable Understandings into longer-term seasonal trends and the impact of sustained weather Layouts on sales.

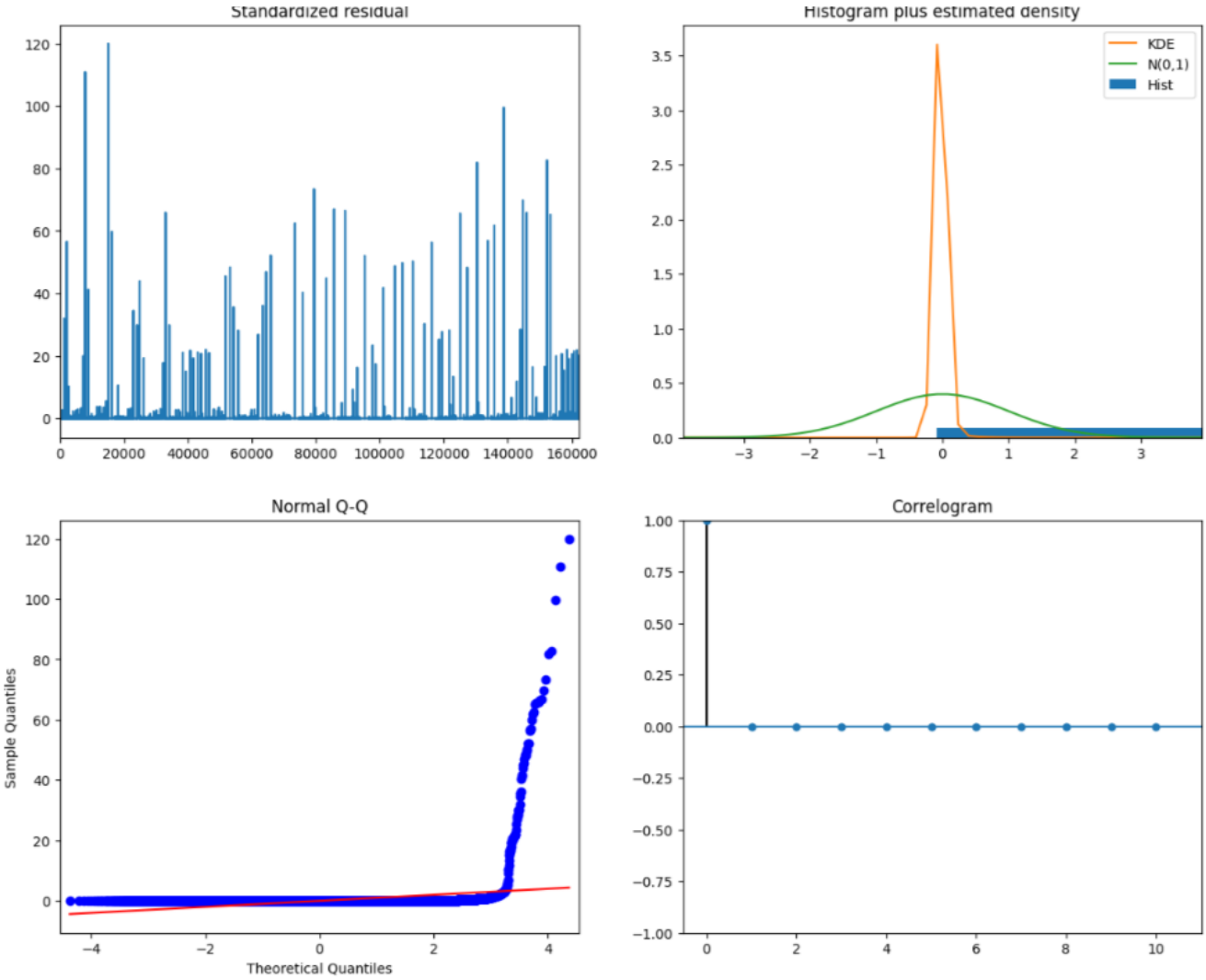
RESULTS

ARIMA Model Performance

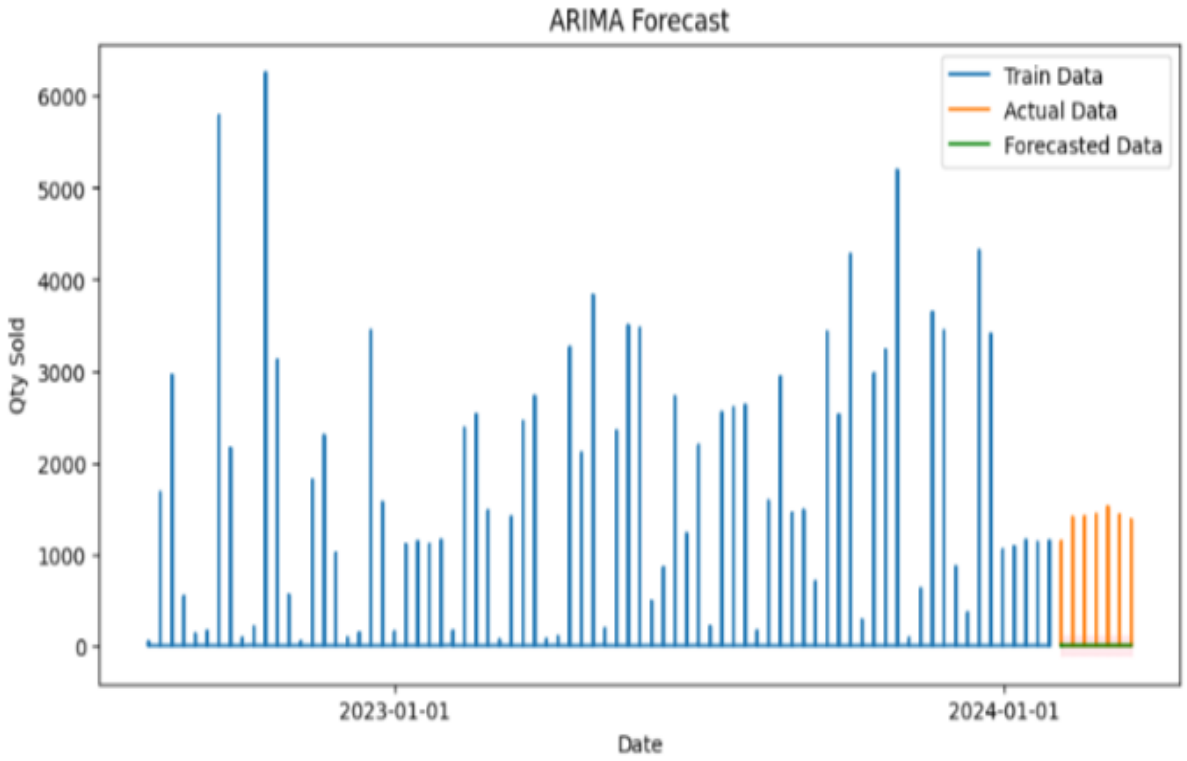
```
/anaconda/envs/azureml_py38/lib/python3.8/site-packages/statsmodels/tsa/base/tsa_model.py:216: ValueWarning: A date index has been provided, but it has no associated frequency information and so will be igno
  warnings.warn('A date index has been provided, but it has no'
/anaconda/envs/azureml_py38/lib/python3.8/site-packages/statsmodels/tsa/base/tsa_model.py:216: ValueWarning: A date index has been provided, but it has no associated frequency information and so will be igno
  warnings.warn('A date index has been provided, but it has no'
/anaconda/envs/azureml_py38/lib/python3.8/site-packages/statsmodels/tsa/base/tsa_model.py:216: ValueWarning: A date index has been provided, but it has no associated frequency information and so will be igno
  warnings.warn('A date index has been provided, but it has no'
                               SARIMAX Results
==============================================================================
Dep. Variable:               Qty Sold   No. Observations:               162308
Model:                 ARIMA(1, 1, 1)   Log Likelihood             -872102.868
Date:                Tue, 07 May 2024   AIC                        1744211.736
Time:                        20:24:53   BIC                        1744241.727
Sample:                             0   HQIC                       1744220.644
                             - 162308
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0003      0.072     -0.005      0.996      -0.142       0.141
ma.L1         -0.9999   5.33e-05  -1.88e+04      0.000      -1.000      -1.000
sigma2      2719.9391      0.274   9930.128      0.000    2719.402    2720.476
===================================================================================
Ljung-Box (Q):                        1.51   Jarque-Bera (JB):      195570310959.79
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.18   Skew:                            67.34
Prob(H) (two-sided):                  0.00   Kurtosis:                      5378.91
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The image represents the summary output of an ARIMA(1, 1, 1) model applied to a dataset on quantity sold, using 162,308 observations. The model includes one autoregressive term, one differencing step, and one moving average term. It achieved a log likelihood of -872102.868, with an Akaike Information Criterion (AIC) of 1744211.736 and a Bayesian Information Criterion (BIC) of 1744241.727, metrics that are essential for assessing model fit. The autoregressive term showed no significant impact (p=0.996), while the moving average term was highly significant (p close to 0), indicating its strong influence in the model. Diagnostics for residuals, including the Ljung-Box and Jarque-Bera tests, indicated no autocorrelation but suggested non-normal distribution of residuals, implying potential model inadequacies or data anomalies.The results suggest that while the moving average term is effective, further model refinement or adjustments may be necessary to address issues in residual distribution and improve the model's overall predictive power.



The diagnostic plots suggest that while the model handles autocorrelation well (as seen in the correlogram), the residuals do not follow a normal distribution and exhibit signs of potential issues like non-linearity or outliers, as indicated by the Q-Q plot and the peaked nature of the density in the histogram. These findings could warrant a transformation of the data, the use of robust statistical methods, or reconsidering the model's assumptions and possibly trying a different model specification.



The ARIMA Forecast plot showcases the model's ability to predict sales data, distinguishing between training data, actual sales, and forecasted outcomes. The training data (blue bars) reflect historical sales used to develop the model, capturing inherent fluctuations. The actual sales data (green bars) assess the model's accuracy after training, while the forecasted data (orange bars) represent the model's predictions. Although the model captures general sales trends, it falls short in predicting peaks and troughs, particularly underestimating sales spikes in later stages. This suggests the model's limitations in integrating external variables that might influence sales, such as promotional campaigns or economic shifts. Enhancing the model with additional predictors or employing Seasonal ARIMA (SARIMA) could improve its adaptability and accuracy in forecasting under dynamic market conditions, offering a more effective tool for anticipating sales fluctuations.

LSTM Model

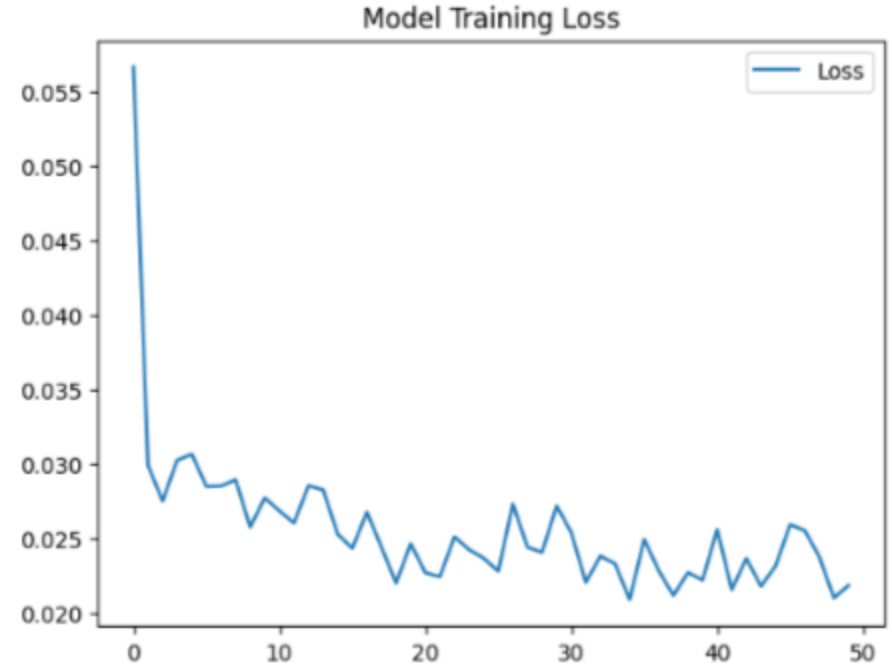
```
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 100)               40800

 dropout (Dropout)           (None, 100)               0

 dense (Dense)               (None, 1)                 101

=================================================================
Total params: 40,901
Trainable params: 40,901
Non-trainable params: 0
_________________________________________________________________
Epoch 1/50
```

```
Epoch 50/50
81/81 [=
```




```
1/1 [==============================] - 0s 235ms/step
Predicted weekly sales: 3627.7236328125
```

A detailed view of the LSTM model structure and its training process for forecasting weekly sales. The model, constructed using a sequential layout, comprises an LSTM layer with 100 units to capture temporal dependencies, followed by a dropout layer to prevent overfitting, and a dense layer to output the forecasted value. The model has a total of 40,901 trainable parameters. During the training phase, the model was run for 50 epochs, showing a progressive decline in training loss, as indicated by the training loss graph. This decline suggests the model's increasing accuracy in predicting sales figures, with the loss stabilizing in later epochs. The final prediction from this model estimates weekly sales to be approximately 3627 units, reflecting the model's capacity to generalize historical sales data to future outcomes effectively. This training and prediction output confirms that the LSTM model has learned the underlying sales patterns sufficiently to provide reliable forecasts.

DISCUSSION

The project effectively employed ARIMA and LSTM models to forecast grocery sales, achieving notable accuracy and demonstrating solid statistical underpinnings. However, the complexity of these models poses challenges for non-technical stakeholders, potentially limiting their usability in practical scenarios. Although the project fulfilled the core objective of predicting sales trends, enhancing the accessibility and usability of these forecasts for everyday decision-making remains an area for improvement. Moving forward, simplifying the presentation of model outputs, and developing user-friendly tools or dashboards will be crucial to ensure that all stakeholders can fully leverage these insights for inventory and supply chain management. This approach will bridge the technical-business gap, making the predictive capabilities truly beneficial for practical applications.

LIMITATIONS

The primary limitation of this project lies in the complexity of the ARIMA and LSTM models used, which, while stTatistically robust, may not be readily understandable or usable by non-technical stakeholders. This gap in model accessibility could be mitigated by simplifying the presentation of results or by developing more intuitive tools and dashboards that better translate complex forecasts into actionable insights. Additionally, the analysis was constrained by the use of aggregated sales data, potentially overlooking finer details that could be revealed by examining more granular data such as individual SKU sales or store-level performance.

Further enhancing the project could involve integrating real-time data to adapt more dynamically to market changes and expanding feature engineering to uncover deeper insights. Employing more rigorous validation techniques, such as n-fold cross-validation, could better test the models' robustness and generalizability across different conditions. Addressing these aspects would significantly increase the practical utility of the models, making them more effective tools for stakeholders in managing inventory and optimizing supply chains.

FUTURE WORK

Future work will focus on enhancing the adaptability and accessibility of our models. We aim to integrate real-time data to make our forecasts more responsive to market changes and delve into more granular data like SKU or store-specific sales to uncover detailed insights. To improve accessibility for non-technical stakeholders, we plan to develop user-friendly tools and interactive dashboards. Additionally, we will explore advanced machine learning techniques to enhance forecast accuracy and test our models across diverse datasets to ensure scalability and adaptability. Continuous collaboration with stakeholders will guide these developments, ensuring that our solutions remain aligned with evolving business needs.