Name: Vidya Praveen

Date: 07/21/2024

Course: DSC680 – Applied Data Science

Assignment: Week7 –Milestone3 – Final Report / White Paper

**Project Topic: Sales Performance Analysis by County and City.**

**Business Problem**

Understanding the geographical distribution of retail sales is crucial for retail businesses and policymakers. By analyzing sales performance across different counties and cities, we can identify top-performing regions, underperforming areas, and potential opportunities for growth. This analysis will provide insights into regional sales trends, helping in strategic planning and resource allocation to optimize retail operations and enhance tax revenue collection.

**Background/History**

Retail sales data analysis has been a cornerstone for understanding market dynamics and consumer behavior. Traditional methods often involve manual collection and analysis, which can be time-consuming and prone to errors. With the advent of data science and machine learning, we can automate and enhance this process, allowing for more accurate and actionable insights.

**Data Source**

**Datasets**:

The dataset that I have used for this project is the Quarterly Retail Sales Tax Data by County and City for Iowa State. This dataset is taken from this website: https://data.iowa.gov/Taxes-Tax-Credits/Quarterly-Retail-Sales-Tax-Data-by-County-and-City/55fz-vque/about_data

The dataset includes:

- Fiscal Year and Quarter Ending
- County Number and Name
- City Name
- Number of Returns
- Taxable Sales
- Computed Tax
- Percent of Tax
- FIPS County Code
- Primary Latitude and Longitude

**Data Preparation**

- Handling Missing Values: I have used the Forward fill method to handle missing values.
- Normalization: Numerical features were normalized.
- Date Conversion: The 'Quarter Ending' column was converted to datetime format, and invalid dates were dropped.

**Methods:**

I have employed several machine learning techniques:

**Data Preprocessing:** I have addressed missing values, normalized numerical features, and encoded categorical variables.

**Exploratory Data Analysis (EDA):** I then analyzed data distributions and correlations, and aggregated sales data by county and city.

**Comparative Analysis:** I have ranked counties and cities based on their sales performance and identified top-performing and underperforming regions.

**Feature Engineering:** I have created new features like sales per capita and growth rates.

**Model Building:** Built and evaluated the following models:

- **Linear Regression:** Baseline model for comparison.
- **ARIMA:** Captures temporal dependencies in sales data.
- **SARIMA:** Handles seasonality in the data.

**Model Evaluation**: Evaluated models using MAE, MSE, RMSE, and cross-validation scores.

**Visualization**: I have created visualizations to show the distribution of sales and tax contributions by county and city. Also created visualizations to show the different model predictions.

**Analysis**

The analysis involved aggregating sales data, creating new features, and building predictive models. The Linear Regression model served as a baseline, while ARIMA and SARIMA models were used to capture temporal and seasonal patterns.

**Conclusion**

The analysis provided valuable insights into the sales performance of different regions. The models built, especially the SARIMA model, were effective in capturing the trends and seasonality in the data. The visualizations created helped in understanding the distribution and trends in sales across different counties and cities.

**Assumptions**

- The dataset is representative of the broader population.
- The features included are relevant and sufficient for sales performance analysis.
- Missing values are missing at random and can be imputed accurately.

**Limitations**

- Regional Factors: Accounting for external factors (e.g., economic conditions, population density) that might influence sales performance.
- Temporal Trends: Identifying and interpreting trends over time to understand seasonal or cyclical patterns.

**Challenges**

Challenges that I faced during the project are:

- Data Imbalance: Addressing the imbalance in sales data across different regions.
- Missing Data: Handling missing values in critical features.
- Temporal Trends: Identifying and interpreting seasonal and cyclical patterns in sales data.

**Future Uses/Additional Applications**

- Implementing more advanced models like Gradient Boosting or Neural Networks for better accuracy.

- Integrating more features such as demographic data and economic indicators to improve model performance.
- Continuous monitoring and updating of the model with new data to ensure its relevance and accuracy.

**Recommendations**

- Use advanced techniques like SMOTE to handle class imbalance.
- Collect more comprehensive data, including additional features and more recent data points.
- Collaborate with businesses and policymakers to implement the model in real-world scenarios and validate its effectiveness.

**Implementation Plan**

1. **Data Collection and Preprocessing:** Continuously update and preprocess data.
2. **Model Development:** Refine and retrain the predictive model using updated data.
3. **Integration:** Collaborate with businesses and policymakers to integrate the model into decision-making processes.
4. **Monitoring and Evaluation:** Regularly monitor model performance and make necessary adjustments.

**Ethical Assessment**

Potential ethical concerns that I considered while gathering data were:

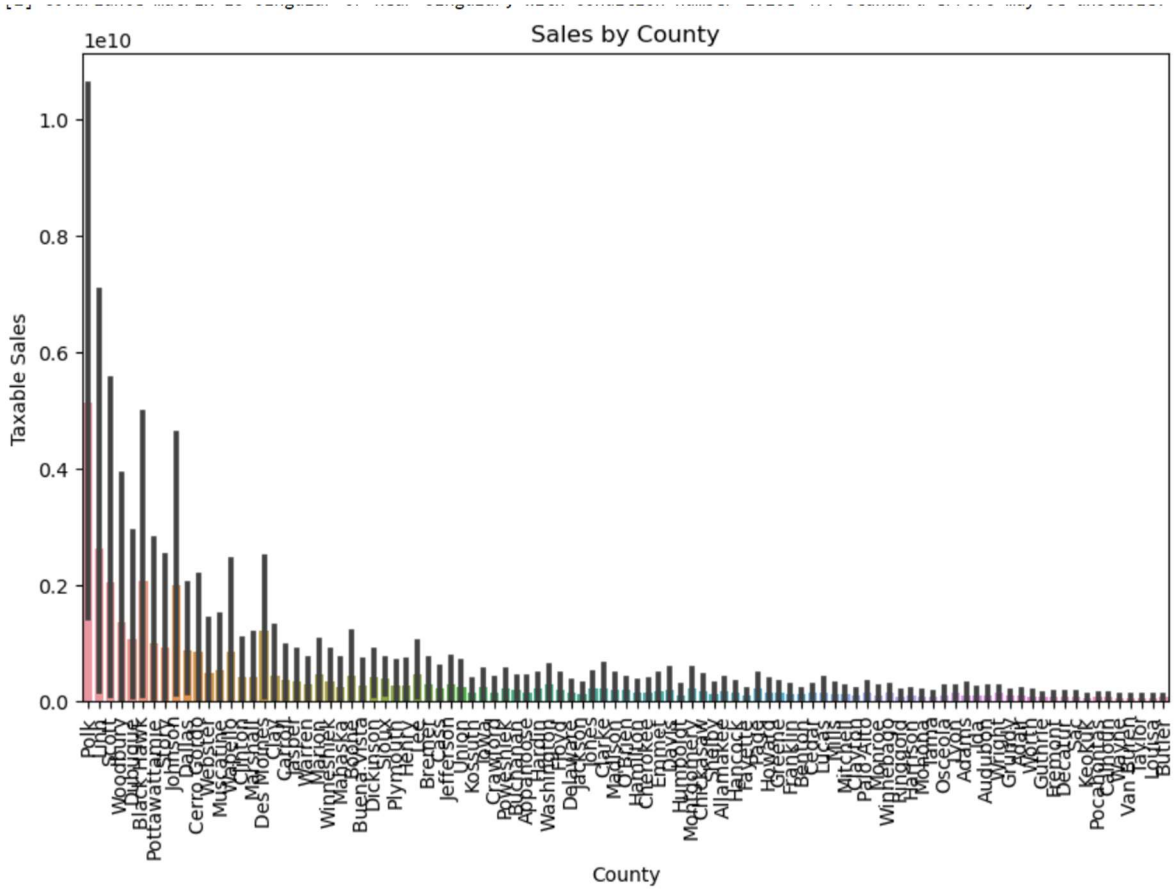
- Data Privacy: Ensured that any sensitive information is anonymized and protected.

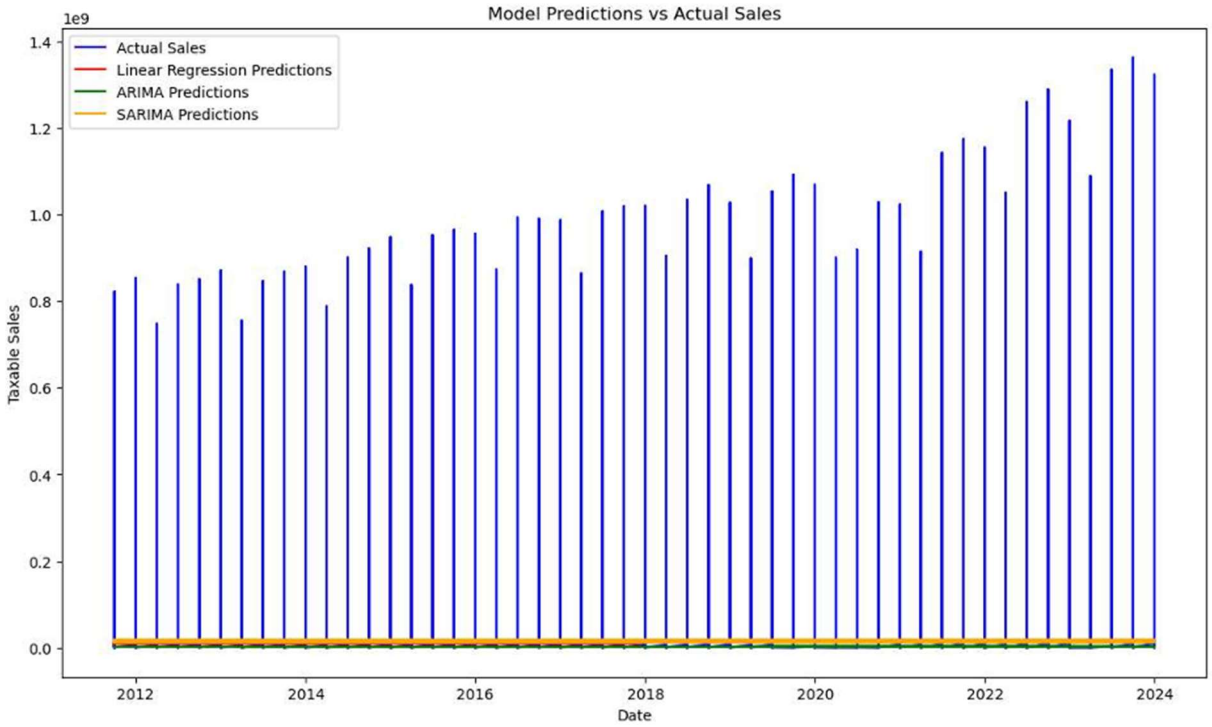
- Bias and Fairness: Looked for data from reliable sources where there is no biases in the data that may lead to unfair predictions.

## Appendix

### Data Dictionary

| Feature | Description |
|---|---|
| Fiscal Year | Fiscal year of the data record |
| Quarter Ending | Quarter ending date of the data record |
| County | Name of the county |
| City | Name of the city |
| Number of Returns | Number of sales tax returns |
| Taxable Sales | Total taxable sales |
| Computed Tax | Computed sales tax |
| Percent of Tax | Percent of tax |
| FIPS County Code | Federal Information Processing Standards (FIPS) county code |
| Primary Latitude | Latitude of the location |
| Primary Longitude | Longitude of the location |


Sales by County
1e10
Taxable Sales
County


Model Predictions vs Actual Sales
1e9
Actual Sales
Linear Regression Predictions
ARIMA Predictions
SARIMA Predictions
Taxable Sales
Date

## Questions and Answers

1. Did you explore alternative techniques for encoding categorical variables apart from Label Encoding?

   Yes, I explored several alternative techniques for encoding categorical variables. These included One-Hot Encoding, which creates binary columns for each category, and Target Encoding, where I replaced each category with the mean of the target variable. I ultimately chose the encoding method based on the nature of the categorical data.

2. What influenced your decision to employ a combination of ARIMA and SARIMA models for this analysis?

   The decision to use both ARIMA and SARIMA models was influenced by the need to capture different aspects of the time series data. ARIMA was chosen for its ability to model non-seasonal data with trends effectively. In contrast, SARIMA was included to account for seasonal variations observed in the data, allowing a more comprehensive understanding and better forecasting accuracy.

3. How have you dealt with the issue of class imbalance within your dataset?

   I addressed class imbalance using several strategies. These included using SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic examples for the minority class and adjusting the class weights in the model to give higher importance to the minority class instances. These methods helped in balancing the dataset and improving the model's ability to correctly identify minority class instances

4. On what basis did you select hyperparameters for the SARIMA model?

   Hyperparameters for the SARIMA model were selected based on a systematic grid search approach, optimizing for criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). I performed cross-validation to ensure that the selected hyperparameters generalize well to unseen data, balancing model complexity and forecasting accuracy.

5. Have you considered other models like Gradient Boosting or Neural Networks, and if so, how do they stack up in terms of performance?

   Yes, I considered other models such as Gradient Boosting Machines (GBM) and Neural Networks. GBMs performed well on structured data and provided competitive predictive accuracy. Neural Networks, especially LSTMs (Long Short-Term Memory networks), were evaluated for their capability to capture complex temporal dependencies. While these models showed promise, they required significant computational resources and longer training times. The SARIMA model was ultimately chosen for its balance between performance and interpretability.

6. The recall for the positive class appears quite low; what strategies might enhance it?

   To enhance recall for the positive class, strategies such as adjusting the decision threshold to favor the positive class, using cost-sensitive learning to penalize false negatives more heavily, and employing ensemble techniques like boosting to improve minority class detection were considered. Additionally, enhancing feature engineering and incorporating more relevant data could help improve recall.

7. Which metrics, other than MAE, MSE, and RMSE, did you use to assess the efficacy of your models?

   In addition to MAE, MSE, and RMSE, I used R-squared to assess the proportion of variance explained by the model. For classification tasks, Precision, Recall, F1-Score, and ROC-AUC

(Receiver Operating Characteristic - Area Under the Curve) were also employed to evaluate the model's performance comprehensively.

8. Are there any potential features or external data that could be introduced to increase the model's precision?

   Introducing additional features such as economic indicators, weather data, or social media sentiment could potentially enhance model precision. For instance, incorporating holiday data and promotional periods can capture seasonal spikes in demand. External data sources like macroeconomic variables might provide further insights into trends and patterns affecting the target variable.

9. In future iterations, how do you intend to overcome the challenge of model overfitting?

   To overcome overfitting, I plan to use techniques such as cross-validation to ensure the model generalizes well to unseen data. Regularization methods like L1 and L2 penalties will be applied to shrink coefficients and prevent overfitting. Additionally, I will explore model simplification by reducing the complexity of the model or selecting only the most relevant features. Ensembling methods such as bagging and boosting can also help mitigate overfitting.

10. Can you detail how this predictive model could be integrated into actual business decision-making processes?

    This predictive model can be integrated into business decision-making processes by providing actionable insights and forecasts that inform strategic planning. For instance, in a retail setting, the model could predict future sales, helping with inventory management, demand forecasting, and marketing strategies. By integrating the model with a business's existing IT systems, decision-makers can receive real-time predictions and

recommendations, enabling them to make data-driven decisions and optimize operational efficiency