Shreya Bhardwaj sb2182 (section 8)
Vaishnavi Nibhanupudi vsn11 (section 7)

Project Definition:

Our project aims to accurately forecast Toyota car sales in New Jersey. This involves predicting how many units will be sold based on various factors such as price, promotions, seasonality, regional economic indicators, and competitor behavior. Reliable sales forecasting can help dealerships and manufacturers optimize inventory, reduce overstock costs, and ensure timely availability of popular models to meet customer demand.

In the real world, the results of this project can address strategic aspects such as inventory management, pricing and promotions strategies, resource allocations, and risk mitigation. By using predictive modeling to anticipate future sales, car companies and dealerships can better allocate and order resources. Analyzing the effects of pricing and promotion on overall car sales can lead to the development of more effective pricing strategies. Also, using this project to analyze the effects of promotions on overall sales can aid in marketing efforts. Finally, the inclusion of economic factors in our project provides valuable insight into how realworld market situations, as well as the economic situations of potential buyers, affects the quantity of sales.

Our project's content reflects the data science and machine learning concepts we covered in class including data management, exploratory analysis, and machine learning modeling. We managed data through creating a dataset, handling any missing values, and performing SQL queries. Then we conducted exploratory data analysis to observe the distributions/correlations between factors to aid our feature selection. Finally, we applied predictive machine learning models such as linear regression and random forest.

Novelty & Importance:

This project is important in understanding car sales trends and is essential for dealerships to remain competitive in today's dynamic market. Fluctuations in economic conditions, customer price sensitivity, and regional preferences require dealerships to make informed, data-driven decisions regarding pricing, inventory, and promotions. This project focuses on identifying key factors, such as seasonal demand, trade-in discounts, and economic indicators, that have the most significant impact on sales. Using these insights, dealerships can optimize their strategies, allocate resources effectively, and align their operations with customer needs.

This project excites us because it connects directly to real-world applications, offering insights into car sales trends and consumer behavior. Understanding the factors that drive car sales provides a practical challenge with meaningful outcomes. Additionally, working with this dataset allows us to apply the skills and concepts we've learned in class, like predictive modeling and data visualization, to a relevant and engaging topic.

A major issue in current data management is the lack of proper integration and understanding of complex data. Many businesses rely on outdated or incomplete datasets, which can lead to biased or narrow insights. Additionally, key factors such as seasonal trends, economic conditions, or regional variations are often overlooked, resulting in overly simplified models. Another common problem is making decisions based on intuition rather than data-driven evidence, which limits the potential for growth and optimization. Tackling these gaps with robust cleaning, feature selection, and thoughtful modeling ensures more accurate predictions and actionable insights.

Progress & Contribution:

We generated a fully synthetic dataset that simulates Toyota car sales in New Jersey. We struggled to find accessible and comprehensive real-world data, so we carefully designed a data generation process to simulate realistic records. To ensure that our data reflects unpredictable real world scenarios and wasn't too "perfect," we added missing values and handled these using imputation techniques. During data generation, we incorporated seasonal patterns, external shocks, and noise. Below are the factors we considered when simulating our data.

1. **Regions and Economic Tiers:**
    - The simulated data categorizes sales by three regions in New Jersey: North Jersey, Central Jersey, and South Jersey.
    - Each region is associated with an economic tier:
        - **North Jersey:** Wealthiest region → higher price factors.
        - **Central Jersey:** Middle-class region → average price factors.
        - **South Jersey:** Lower-income region → lower price factors.
    - We incorporated regional probabilities and price adjustments to mimic how economic status impacts car sales and used np.random.choice to accomplish this
2. **Car Models and Attributes:**
    - We incorporated eight Toyota car models, with varying attributes like:
        - Fuel efficiency (MPG).
        - Powertrain type (Gasoline or Hybrid).
        - Safety ratings.
        - Manufacturer Suggested Retail Price (MSRP).
    - These attributes influence customer preferences and pricing variability.
    - We gave each model a certain popularity value to account for the diverse demand.
3. **Pricing Variability:**
    - Base prices were adjusted using:
        - Regional price factors.
        - Random price variations to mimic dealership-level adjustments. (we used np.random.uniform)
        - Seasonal multipliers to account for demand fluctuations during different times of the year.
        - We applied a variation of +/- 15% to each MSRP of a model
4. **Promotions:**
    - We included seasonal and region-specific promotion probabilities.
    - We represented promotions as discount percentages, ranging from 0% to 20%, and adjusted the likelihoods of promotions by season and region.
5. **Trade-in Discounts:**

  - We included varying discounts based on the car's price, mimicking customer incentives for trading in older vehicles:
    - Higher discounts for more expensive models.
    - Lower discounts for entry-level models.

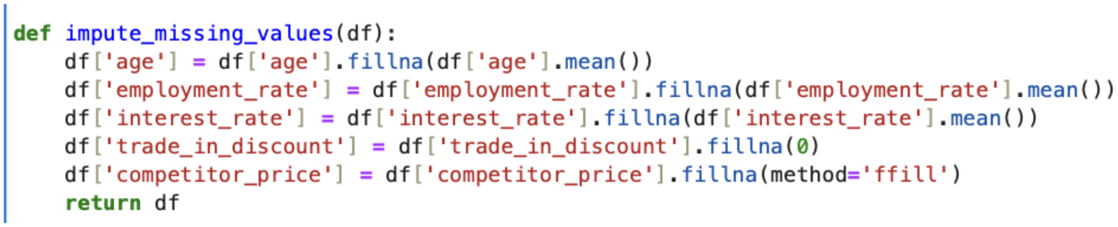
6. **Seasonality:**
   - Seasonal multipliers affected demand:
     - **Winter:** Lower sales due to harsh weather.
     - **Summer:** Higher sales due to increased travel and outdoor activities.
     - We added random noise to ensure realistic variability.
7. **Interest Rates and Employment Rates:**
   - We captured fluctuations in interest rates by establishing a base rate and then added noise to account for the ups and downs of interest rates (this was done through the np.sine manipulation)
   - We established a base value for employment rates and then took account for the inverse relationship it has with interest rates.
   - In each case we used np.random.normal(x,y,z) to add variation to our rates (to reflect real-world market trends) with x representing the centered value and y representing the standard deviation
   - We also included randomness to reflect external shocks.
8. **Competitor Prices and Shocks:**
   - We included competitor prices with monthly shocks and randomness to reflect real world market competition.
   - We then compared final prices and competitor prices
9. **External Shocks:**
   - We included external shocks like supply chain disruptions and fuel price changes to affect sales volumes.
10. **Customer Demographics:**
    - We made three categories of age groups (young, middle aged, and older) and randomly generated customer age.
11. **Quantity Sold:**
    - We calculated Sales quantities based on:
      - Base sales volumes.
      - Pricing factors (inverse relationship with price).
      - Promotion boosts.
      - Seasonal and regional factors.
      - External shocks, adding randomness to reflect unpredictable events.

Models, Techniques, and Algorithms Used:

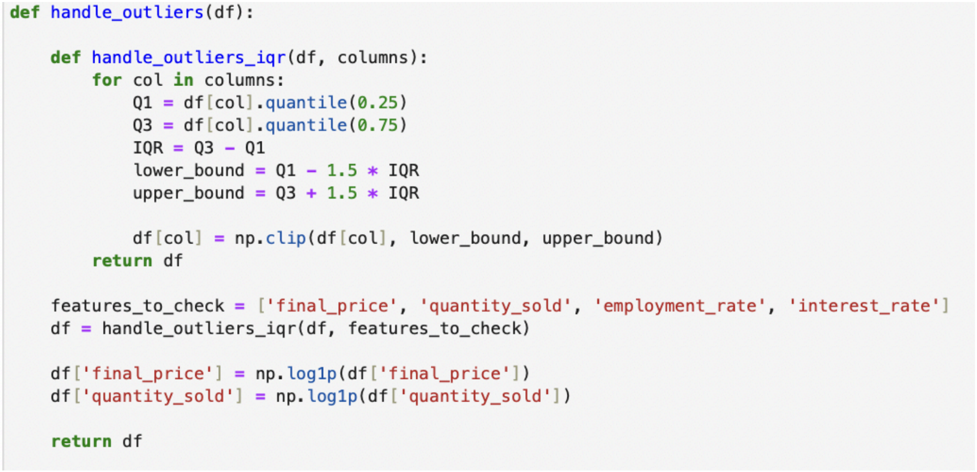
Data Handling Techniques:

1. **Feature Engineering and Selection**:

- We adjusted and refined features to reduce redundancy and improve model interpretability. When there were redundant features, we simplified these by ensuring that predictors contributed uniquely to the prediction target and didn’t have dependencies.

2. **Handling Missing Values/Data Cleaning**:
    - We used mean imputation for missing values in interest_rate, employment_rate, and age to reflect the average trends for those factors..
    - We filled in missing trade_in discount values with 0.
    - We used forward fill to fill in competitor discount missing values to keep data consistency.
    - 

```python
def impute_missing_values(df):
    df['age'] = df['age'].fillna(df['age'].mean())
    df['employment_rate'] = df['employment_rate'].fillna(df['employment_rate'].mean())
    df['interest_rate'] = df['interest_rate'].fillna(df['interest_rate'].mean())
    df['trade_in_discount'] = df['trade_in_discount'].fillna(0)
    df['competitor_price'] = df['competitor_price'].fillna(method='ffill')
    return df
```

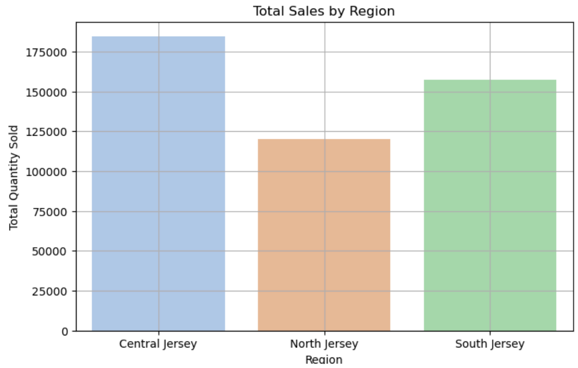
3. **Outlier Detection and Handling**:
    - We handled outliers in the dataset using the Interquartile Range (IQR) method, removing values outside 1.5 times the IQR to the lower and upper bounds for specific columns. We used log to address skewness and stabilize variance which ensured our data to be more consistent and suitable for modeling
    - 

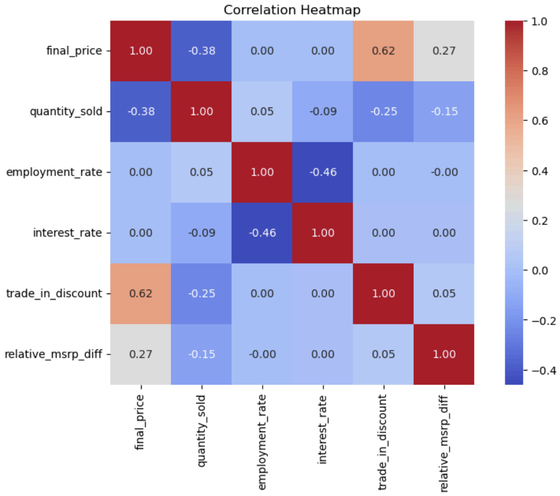
```python
def handle_outliers(df):

    def handle_outliers_iqr(df, columns):
        for col in columns:
            Q1 = df[col].quantile(0.25)
            Q3 = df[col].quantile(0.75)
            IQR = Q3 - Q1
            lower_bound = Q1 - 1.5 * IQR
            upper_bound = Q3 + 1.5 * IQR

            df[col] = np.clip(df[col], lower_bound, upper_bound)
        return df

    features_to_check = ['final_price', 'quantity_sold', 'employment_rate', 'interest_rate']
    df = handle_outliers_iqr(df, features_to_check)

    df['final_price'] = np.log1p(df['final_price'])
    df['quantity_sold'] = np.log1p(df['quantity_sold'])

    return df
```

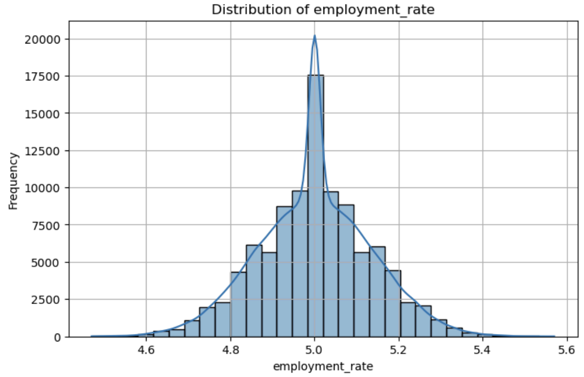
4. **SQL Database Integration**:
    - We inserted the processed dataset into an SQLite database using SQLAlchemy, enabling efficient querying and performed SQL operations like filtering, updating, and aggregating records.
5. **Exploratory Data Analysis:** Analyzing the distributions prompted us to remove outliers as some distributions were skewed or multi-modal, which made outlier handling and imputation critical steps before modeling.

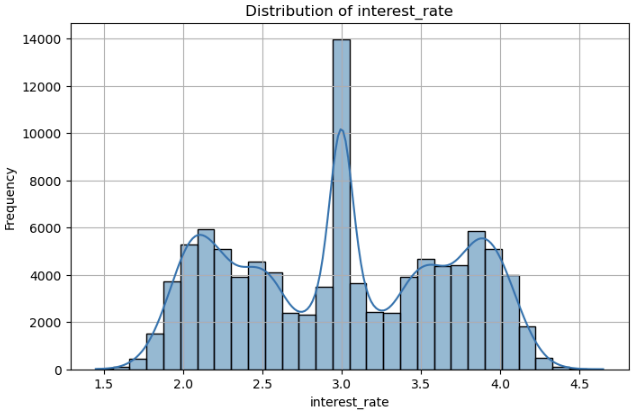
    The interactions between pricing strategies, economic factors, and regional preferences plays a key role in making accurate predictions for car sales. By understanding these connections, models like Random Forest can pick up on complex patterns and highlight the most important factors, making them much better at predicting outcomes.

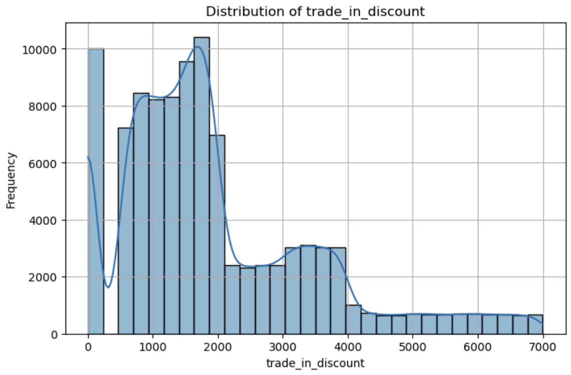
There is a strong positive correlation (0.62) between final price and trade-in discount, suggesting that higher-priced vehicles often receive larger trade-in discounts, while a moderate negative correlation (-0.38) between final price and quantity sold indicates that lower-priced cars drive higher sales volumes. A moderate negative correlation (-0.46) between employment and interest rates aligns with macroeconomic trends, while weak correlations, such as between quantity sold and trade-in discount (-0.25), imply higher trade-in discounts are linked to lower sales volumes for high-end models. Weak positive correlations, like that between final price and MSRP difference (0.27), suggest larger discounts for more expensive cars. The data reflects a price-volume trade-off, where affordability boosts sales, and trade-in discounts are used strategically to sell high-value vehicles. While employment and interest rates don't directly impact sales, they influence purchasing power over time.

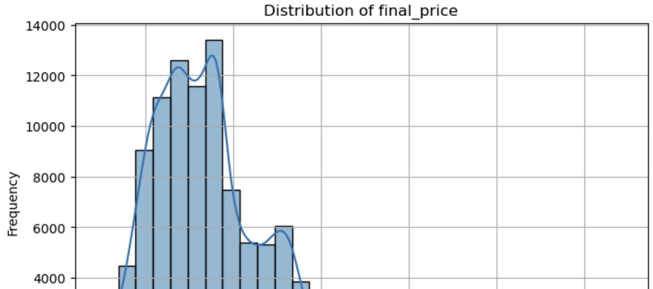


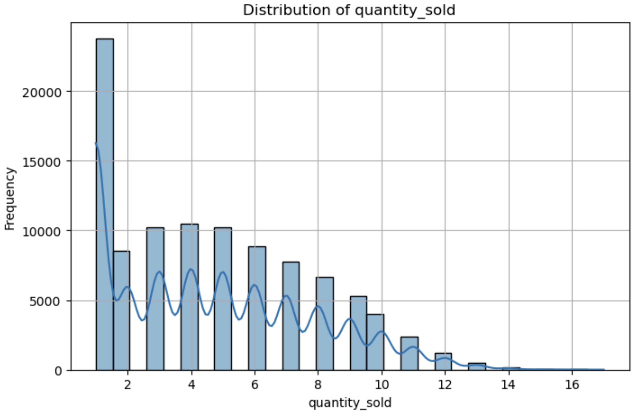












Predictive Model:

Our target or dependent variable was the quantity_sold. We generated and compared the results of two models in predicting this target variable. The independent variables or features included factors such as promotion, employment_rate, interest_rate, region, car_model, final_price, relative_msrp_diff, is_weekend, and season. These predictors provided the model with economic indicators, pricing details, and other factors to help forecast the number of cars sold.

1. **Linear Regression:**
    a. We trained the linear regression model as a baseline model using key predictors such as promotions, employment rate, interest rate, region, car model, and seasonal factors.
    b. We preprocessed the data and encoded categorical variables like region, car model, and season using one-hot encoding.
    c. We derived feature importance model coefficients to identify the most impactful predictors.
    d. We used a scatter plot to visualize the true vs. predicted data.
2. **Random Forest Regressor:**
    a. We used this model to capture nonlinear relationships and obtain higher predictive accuracy if there were more complex relationships in the data.
    b. We used the same key predictors from the linear regression model.
    c. To prevent overfitting we used a maximum depth of 10.

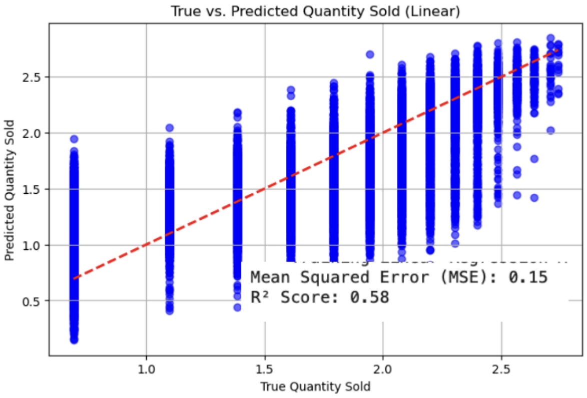
Experiments Designed:

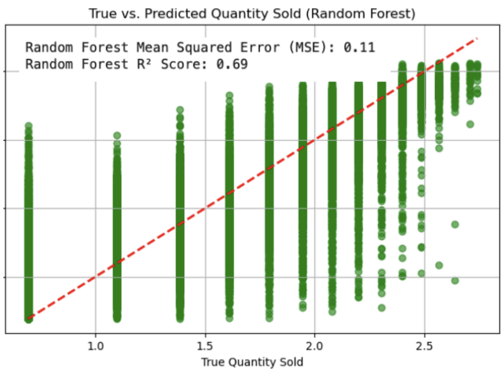
1. We started with a simpler dataset and gradually introduced complexity such as noise, external shocks, seasonal variations, and nonlinearities to see how model performance changed.
2. We tried including or excluding certain features to observe their impact on the model's accuracy and interpretability. In the generation of our final predictive model, we excluded the price feature and used final_price to reduce redundancy.
3. By evaluating both linear regression and random forest on the same dataset, we tested how well each model adapted to increasing complexity and concluded that non-linear methods handled it better.

4. We examined the effects of handling missing values and outliers, comparing performance before and after data cleaning steps.

Key Findings and Results: The project found that promotions, seasonal trends, final price, and regional economic conditions were key factors influencing car sales. Promotions and seasonal multipliers had the strongest positive impact on sales, while higher final prices reduced them. We also found that Random Forest outperformed Linear Regression, achieving lower Mean Squared Error (MSE) and higher R^2 Score, demonstrating that there are non-linear relationships/interactions in the dataset. This makes sense as during the simulation of our data set we did obtain certain features from calculations involving others. Our hypothesis that these factors significantly affect car sales was verified.

Evaluation Method:
We evaluated our approach using performance metrics like Mean Squared Error (MSE) and the R^2 score. As we gradually increased the complexity of the data, we observed these values to ensure that our adjustments made the model more realistic and better at forecasting.





The comparison of true vs. predicted quantity sold shows that the Random Forest model outperformed the Linear Regression model. The Random Forest achieved a lower Mean Squared Error (MSE) of 0.11 and a higher R2R^2R2 score of 0.69, indicating better accuracy and a stronger ability to explain variance in the data. In contrast, the Linear Regression model had a higher MSE of 0.15 and a lower R2R^2R2 score of 0.58, reflecting its limitations in capturing non-linear relationships. The tighter clustering around the red diagonal line in the Random Forest plot demonstrates more accurate predictions compared to the wider spread observed in the Linear Regression plot. This confirms that Random Forest is better suited for handling the complexities and interactions present between features in the dataset.

Advantages:

- The project covers the full workflow for data science: data generation, database integration, exploratory data analysis, model training, and evaluation. This approach showed us how each step contributes to the final predictive mode and provided us with practical skills for data management.
- Using both linear regression and random forest highlights the contrast between simple and more advanced modeling techniques, showing how certain models handle complexity better.

- The inclusion of randomness and external shocks introduces realistic variability, making the dataset better reflect real-world market conditions.

Limitations:

- Despite the randomness we added, the dataset is still artificial and may not capture all the nuances, biases, and unpredictable factors of an actual market.
- We focused on linear regression and random forests, however, more sophisticated models might offer different insights or improved performance.
- Without real-world data, we cannot verify if our results are applicable to the real world.

Changes After Proposal:

Initially, we aimed to use API's to obtain real-world data, but we weren't able to find or access substantial, cohesive data, so we decided to generate our own. Furthermore, we narrowed our scope from all car companies in New Jersey to Toyota so that we could focus on a specific brand, making our simulation and analysis more manageable.