



UNIFIED MENTOR
YOUR SKILL. SUCCESS & JOURNEY

PROJECT – 4

NAME - AAGASH.M

UNID - UMID11062542183

PROJECT TITLE – “ Colorado Motor Vehicle Sales Data”

Abstract

This project provides a detailed analysis of the Colorado motor vehicle market using sales data from 2008 to 2015. The primary goals were to conduct an economic analysis of sales trends, identify market patterns across different counties, and develop a predictive model to forecast future sales. The methodology involved comprehensive exploratory data analysis (EDA) to visualize trends, seasonal patterns, and county-level performance. Subsequently, a Random Forest Regressor model was trained to predict vehicle sales based on year, quarter, and county. The model's performance was evaluated using the Root Mean Squared Error (RMSE). The analysis revealed significant growth in the automotive market post-2009 and identified key counties like Arapahoe and El Paso as major sales drivers. The predictive model demonstrated strong performance, providing a valuable tool for financial analysis, market research, and strategic planning.

Table of Contents

1. Introduction

- 1.1. Background
- 1.2. Problem Statement

2. Objectives

3. Methodology

- 3.1. Dataset Description
- 3.2. Data Preprocessing and Feature Engineering
- 3.3. Exploratory Data Analysis (EDA)
- 3.4. Regression Modeling

4. Results and Discussion

- 4.1. Time-Series and Trend Analysis
- 4.2. County-Level Market Analysis
- 4.3. Predictive Model Performance
- 4.4. Feature Importance

5. Conclusion and Future Work

- 5.1. Conclusion
- 5.2. Future Work

1. Introduction

1.1. Background

Motor vehicle sales are a critical indicator of economic health, reflecting consumer confidence and spending patterns. In a diverse economy like Colorado's, analyzing this data provides valuable insights into regional economic disparities, market trends, and the overall business cycle. This dataset, which contains quarterly sales information for various Colorado counties, serves as a rich resource for financial analysts, market researchers, and policymakers.

1.2. Problem Statement

To make informed business and policy decisions, stakeholders need to understand the dynamics of the automotive market. This requires a systematic analysis of historical sales data to identify trends, forecast future performance,

and determine the key factors driving sales. This project addresses this need by applying data analysis and machine learning techniques to the Colorado motor vehicle sales data.

2. Objectives

The primary objectives of this project are:

- **Economic & Trend Analysis:** To analyze and visualize historical sales data to identify long-term trends, seasonal patterns, and significant economic events.
- **Market Research:** To compare sales performance across different Colorado counties to identify key markets and regional disparities.
- **Sales Forecasting:** To build and evaluate a machine learning regression model capable of accurately predicting future vehicle sales.
- **Feature Importance:** To identify the most significant drivers of vehicle sales from the available data (e.g., year, quarter, specific counties).

3. Methodology

3.1. Dataset Description

The dataset contains motor vehicle sales records for multiple counties in Colorado, recorded quarterly from 2008 to 2015.

Key Columns:

- **year:** The calendar year of the sales record.
- **quarter:** The quarter of the year (1-4).
- **county:** The name of the Colorado county.
- **sales:** The total dollar amount of motor vehicle sales.

3.2. Data Preprocessing and Feature Engineering

1. **Date Feature Creation:** A date column was created by combining the year and quarter to facilitate time-series plotting.
2. **Handling Categorical Data:** The county column, being categorical, was transformed using one-hot encoding. This process converts each county name into a separate numerical column, making it suitable for the machine learning model.

3.3. Exploratory Data Analysis (EDA)

EDA was performed to uncover insights visually. This included plotting total sales over time to observe trends, aggregating sales by county to identify top markets, and analyzing average sales per quarter to find seasonal patterns.

3.4. Regression Modeling

A **Random Forest Regressor** was chosen for the forecasting task. This model is an ensemble of decision trees and is well-suited for this type of tabular data, as it can capture non-linear relationships and is robust to overfitting. The data was split into training (80%) and testing (20%) sets. The model was trained on the training data to predict the sales value.

4. Results and Discussion

4.1. Time-Series and Trend Analysis

The time-series plot of total sales shows a distinct "V-shape" pattern, with a sharp decline corresponding to the 2008-2009 financial crisis, followed by a steady and strong recovery in subsequent years. This highlights the automotive market's sensitivity to macroeconomic conditions.

4.2. County-Level Market Analysis

Arapahoe County consistently registered the highest sales volume, followed by other major metropolitan counties like El Paso, Adams, and Jefferson. This indicates that the state's vehicle market is heavily concentrated in its urban centers.

4.3. Predictive Model Performance

The Random Forest model was evaluated on the unseen test data. The performance was measured using the **Root Mean Squared Error (RMSE)**, which indicates the average magnitude of the prediction error in dollars. The model achieved a low RMSE relative to the scale of the sales figures, demonstrating its effectiveness in forecasting sales.

4.4. Feature Importance

The model identified the year as the most important feature, confirming the strong time-based trend observed in the EDA. Following the year, specific counties like Arapahoe, El Paso, and Adams were the next most influential factors, reinforcing their market dominance.

5. Conclusion and Future Work

5.1. Conclusion

This project successfully analyzed the Colorado motor vehicle sales data, uncovering key economic trends and market drivers. The Random Forest Regressor proved to be an effective tool for forecasting sales, providing actionable insights for business planning and financial analysis. The findings confirm that the market is heavily influenced by time-dependent economic trends and is geographically concentrated in major counties.

5.2. Future Work

- **Incorporate External Data:** Enhance the model by adding macroeconomic indicators like unemployment rates, interest rates, or state GDP.
- **Use Advanced Models:** Explore more sophisticated time-series forecasting models like ARIMA or LSTM for potentially higher accuracy.
- **Expand Dataset:** Incorporate more recent data (post-2015) to analyze the market's evolution in the last decade.

DATA SET -

<https://drive.google.com/file/d/18nYH5vic8k7tCYLo0YiXlbVs4YTAA3Tr/view?usp=sharing>

DATA VISUALIZATION COLAB LINK

https://colab.research.google.com/drive/1kUaylygPJ_uS37o1nRWzIf1VE-a64H7B?usp=sharing