Store Sales - Time Series Forecasting

Group: PSW --- ID: B4

Introduction

The retail industry faces challenges with outdated forecasting methods. This project aims to revolutionize this with machine learning. Here's the quick rundown:

Implications:

- Better inventory management for retailers
- Improved product availability for consumers
- Economic benefits through efficient resource use

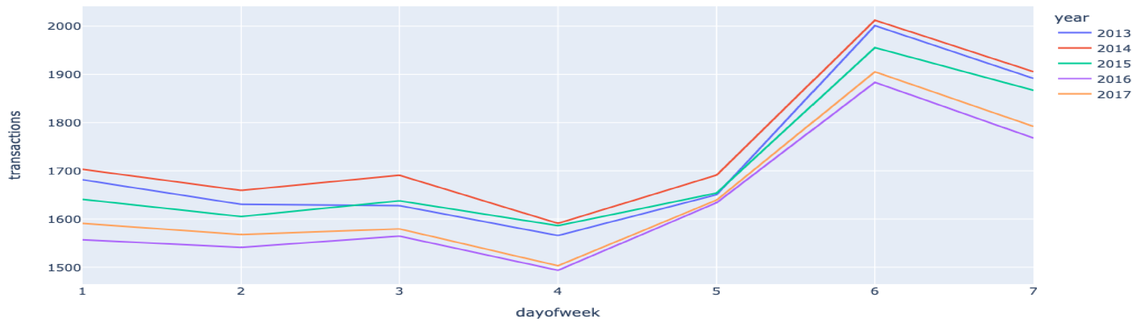
Goals and Hypothesis:

- Develop a machine learning model for accurate retail demand forecasting
- Hypothesize that this model will surpass traditional methods in dynamic retail settings

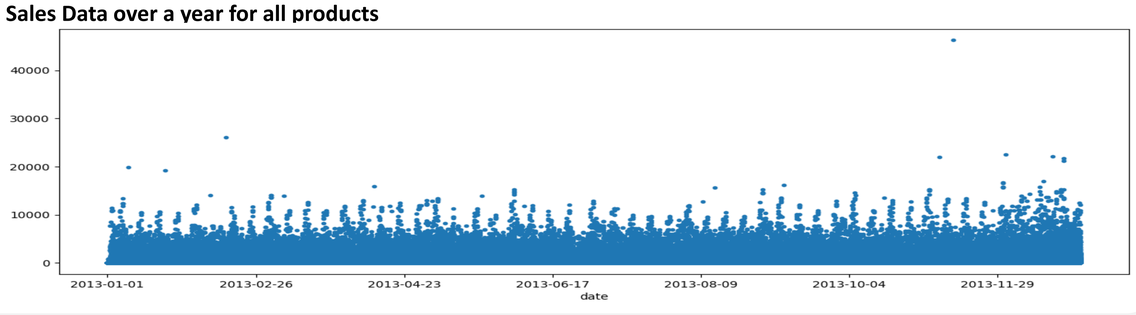
This work isn't just technical; it's about reshaping retail to meet customer needs more effectively. Our poster highlights these key points, underscoring the real-world impact of this machine learning innovation.

Photo of a grocery store in Ecuador

Total Transactions by Day



Sales Data over a year for all products



Data Exploration

The following files were used to engineer useful features and train the models:

- Train.csv – wealth of data relating to the type of product sold, how much of it was sold, how many items are on sale, and what store sold it
- Stores.csv – information about the individual stores and what stores are similar
- Oil.csv – daily oil price (Ecuador is oil dependent)
- Holidays.csv – details about Ecuadorian holidays

Feature Engineering

We refined feature engineering for efficiency. We added lag and rolling window features to 'onpromotion' for capturing short-term trends. Rate of change calculations and handling of missing values were included for data robustness. Key features like 'days_since_last_payday' and 'is_weekend' were added to reflect consumer behavior patterns. This focused approach to feature engineering aims to bolster our model's accuracy in forecasting product demand.

Results

Kaggle scores are Root Mean Squared Logarithmic Error.

CV scores are regular CV scores.

Based on Kaggle scores our best performing model was **XGBoost**.

Model	Kaggle Score	CV Score
Dummy	3.23779	0.006974
Linear Regression	1.89678	0.552296
Ridge*	1.90115	0.484366
Decision Tree	2.40676	0.717528
XGBoost*	1.8205	2.0300*

Notes

- We've tuned the Ridge and XGBoost models via hyper parameter optimization.
- The XGBoost CV score is RMSLE.

Future Work

- The biggest area for improvement is in feature selection and data refinement. There is always more work to be done in generating new and useful features
- We could also try more models, specifically others made for time series data. This would help a lot as most of the models we tested struggled with time series data specifically

UtahStateUniversity

Link to Data

Link to Code

Carter Watson | *Utah State University*
Spencer Potter | *Utah State University*
Ben Smith | *Utah State University*