Technical Report:

Predictive Machine Learning Algorithms for Analyzing Fuel Price Trends in Toronto

Mahnoor Ahmed

Abstract

Fuel prices in Canada exhibit significant variability influenced by a multitude of factors, including economic conditions, taxation policies, crude oil costs, and market dynamics. This technical report focuses on developing machine learning (ML) models to analyze fuel price trends in Toronto using historical data. The objective is to evaluate the performance of various ML techniques—Linear Regression (LR), Random Forest (RF), and K-Nearest Neighbors (KNN) regression—in predicting pump prices and identifying key variables driving price fluctuations.

This project is rooted in two core research questions:

1. How accurately can pump fuel prices be predicted using historical data, and which variables influence these changes the most?
2. How do different machine learning models perform in fuel price prediction?

By leveraging the Ontario Fuel Prices dataset (1987–present) and employing tools such as Python and scikit-learn, pandas, Matplotlib and numpy, this report compares the performance of selected ML models using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). The outcomes of this study provide valuable insights into the application of predictive analytics for fuel price forecasting and contribute to enhancing public and private sector decision-making.

Introduction

Fuel price fluctuations affect multiple stakeholders, including consumers, businesses, and policymakers. Predicting these fluctuations is vital for effective financial planning, resource allocation, and economic stability. The primary aim of this research is to analyze historical fuel price data from Toronto, using ML models to uncover trends and evaluate predictive accuracy.

The dataset is sourced from the Ontario Fuel prices archive and filtered to include only Toronto specific data. It contains 454 records and the following variables:

Variable	Description	Mean	Std Dev	Min	Max
Pump Price	Final price at the pump	89.13	35.20	43.10	207.20
Crude Cost	Cost of crude oil	35.54	19.59	10.30	95.30
Federal Excise Tax	Federal tax per liter	9.49	1.10	4.50	10.00
Ontario Tax	Provincial tax per liter	13.75	2.03	8.30	14.70
GST/HST	Federal goods and services tax	8.07	5.29	3.10	23.80
Wholesale Margin	Distributor’s profit margin	14.88	8.45	3.40	44.70
Retail Margin	Retailer’s profit margin	5.92	2.56	-2.30	12.30

The dataset used includes detailed records of pump prices, taxes, crude oil costs, wholesale margins, and retail margins, offering a comprehensive view of factors influencing price variability. This technical report describes the process of cleaning, visualizing, and modeling the dataset and evaluates model performance to determine the most suitable method for forecasting future fuel prices.

Methodology

1. **Data Cleaning**:
    - Focused exclusively on Toronto-specific records
    - Addressed missing values through imputation using column means
    - Standardized numeric variables consistency
2. **Exploratory Data Analysis (EDA)**: Visualizations like box plots, scatter plots, and histograms highlighted key relationships.
    - Crude cost and Wholesale Margin: Strong positive correlation with pump prices
    - Retail margin: exhibits variability but contributes less significantly
3. **Modeling**: Three models were chosen for their predictive capabilities:
    - Linear Regression: A basic yet interpretable model.
    - Random Forest Regression: Captures complex patterns using ensemble learning, with 100 estimators.
    - KNN Regression: A proximity-based method effective for local patterns, with five neighbours
4. **Evaluation Metrics**:
    - Employing MSE, MAE, and R^2 metrics to assess the accuracy and reliability of each model.

Results

Initial evaluations indicate the following:

- **Linear Regression**:

  MSE: 0.004154, MAE: 0.045235, R^2: 0.999997

- **Random Forest Regression**:

  MSE: 1.996031, MAE: 0.889055, R^2: 0.998535

- **KNN Regression**:

  MSE: 12.446062, MAE: 1.878022, R^2: 0.990867

Discussion

The results show that Linear Regression and Random Forest closely algin with actual values, outperforming KNN in accuracy in predicting pump prices, largely due to the simplicity and linear relationships within the dataset. Random Forest follows closely and provides robust predictions while capturing complex, non-linear patterns in the data. KNN regression performs adequately but shows higher errors, suggesting it may not be the optimal choice for this dataset.

Conclusion

These results emphasize that while simpler models like Linear Regression excel in accuracy, advanced models such as Random Forest are valuable for their ability to model complex relationships effectively. The KNN model, despite being accurate in other contexts, may not be well-suited for this application.

This study highlights the importance of using diverse ML models to analyze and forecast fuel prices. The findings underscore the potential of predictive analytics to enhance decision-making processes, particularly in managing price volatility. Future research may explore additional models and expand the dataset to include more granular features, further improving forecasting capabilities.

References:

Al-Dosary, N. M. N., Al-Hamed, S. A., & Aboukarima, A. M. (2019, December 9). *K-nearest neighbors method for prediction of fuel consumption in tractor-Chisel Plow Systems*. Engenharia Agrícola. https://doi.org/10.1590/1809-4430-Eng.Agric.v39n6p729-736/2019

Escribano, A., & Wang, D. (2020). *FORECASTING GASOLINE PRICES WITH MIXED RANDOM FOREST ERROR CORRECTION MODELS.* Working paper. Economics. https://ideas.repec.org/p/cte/werepe/30557.html

Escribano, A., & Wang, D. (2021, August 21). *Mixed Random Forest, cointegration, and forecasting gasoline prices*. International Journal of Forecasting. https://www.sciencedirect.com/science/article/pii/S0169207020301941

Imandoust, S. B., & Bolandraftar, M. (2013). *Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background*. Research Gate. https://www.researchgate.net/publication/304826093_Application_of_K-nearest_neighbor_KNN_approach_for_predicting_economic_events_theoretical_background

Keerthan, J. S., Nagasai, Y., & Shaik, S. (2019, June 8). *Machine Learning Algorithms for Oil Price Prediction*. Research Gate. https://www.researchgate.net/publication/344227822_Machine_Learning_Algorithms_for_Oil_Price_Prediction

Li, Z., Zhou, B., & Hensher, D. A. (2021, October 14). *Forecasting automobile gasoline demand in Australia using machine learning-based regression*. Energy. https://www.sciencedirect.com/science/article/abs/pii/S0360544221025603

Ministry of Energy. (2024b, September 16). Motor fuel prices. Ontario.ca.

https://www.ontario.ca/motor-fuel-prices/

Mouchtaris, D., Sofianos, E., Gogas, P., & Papadimitriou, T. (2021, September 14). *Forecasting natural gas spot prices with machine learning*. MDPI. https://www.mdpi.com/1996-1073/14/18/5782

Sofianos, E., Zaganidis, E., Papadimitriou, T., & Gogas, P. (2024, March 8). *Forecasting East and West Coast gasoline prices with tree-based machine learning algorithms*. MDPI. https://www.mdpi.com/1996-1073/17/6/1296