Future LPG Shipments Forecasting Based on Empty LPG Vessels Data

Hongxia Shi, Jou-Tzu Kao, Rong Liao, Joseph Tsai, Shenyang Yang, Matthew A. Lanham

Purdue University Krannert School of Management, 403 W. State Street, West Lafayette, IN 47907

shi395@purdue.edu; kao21@purdue.edu; rliao@purdue.edu; tsai103@purdue.edu; yang1469@purdue.edu; lanhamm@purdue.edu

Abstract

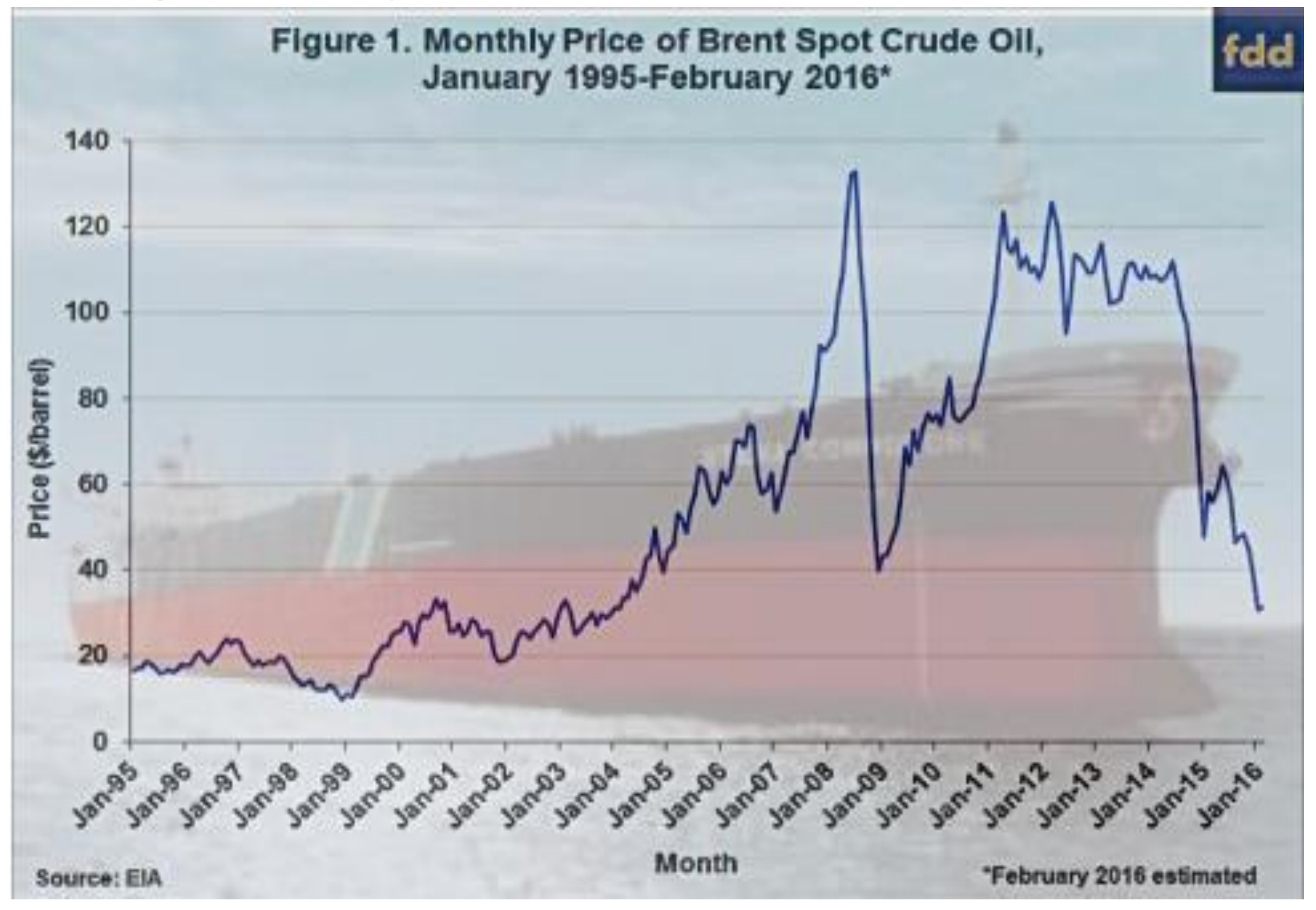
This project assesses the feasibility of using information about empty liquefied petroleum gas (LPG) carrier vessels that are moving within the ocean to predict the how much LPG will be shipped in the future. As prices of multiple commodities fluctuate with the supply and demand of LPG, it is crucial to identify effective indicators of LPG commodity flow to foresee future trends of the market. In order to conduct this analysis, we acquired access to the shipping schedule data of empty and full LPG vessels, and ran multiple types of regressions to understand the correlation between these two factors. After iterative analyses, we obtained a valid ridge regression predictive model among linear, ridge and SVR regressions models, receiving R square of up to 0.8 on test dataset.

Introduction

The primary research question that this study answers is whether the flow of empty shipping movements can serve as an insightful indicator of future LPG shipping. The privilege of tracking and predicting commodity movements will cause phenomenal impact in the business world. Having insight into the flow of goods can allow commodity traders to better predict supply, which provides better decision-support about future price changes.

The motivation of this collaborate project with an industry partner is two fold:

1.Need of predicting future shipments. Whether companies build, buy, or sell commodities, they normally rely on multiple sources of information with great effort to try to make the correct decisions.



Forecast in shipping is especially important in the LPG market, because of the fluctuations of the demand and price as shown above.

2. Need of vessel movements transparency to actionable insights

Maritime transportation is known for having rich information in terms of volume. Most of the transportation information is utilized in supporting the global commodity supply chain. Cooperative self-reporting vessel location systems, including the Automatic Identification System (AIS) and Long-Range Identification and Tracking system (LRIT), provide a great amount information about the vessels at port and at sea. However, converting this geospatial sensor "big data" into better understanding for those in the market is very challenging.

Methodology

Raw Data Sources

The project used the LPG geospatial shipment information available for the full ships, including vessel names, dates, and departing and arriving ports. The feature - the average total empty vessels' capacity is created. The average total empty vessels' capacity is obtained by calculating the empty vessels' capacity from the past 27 to 21 days of a specific date. Figure 2 provides the process we performed from raw data to the complete DSS.

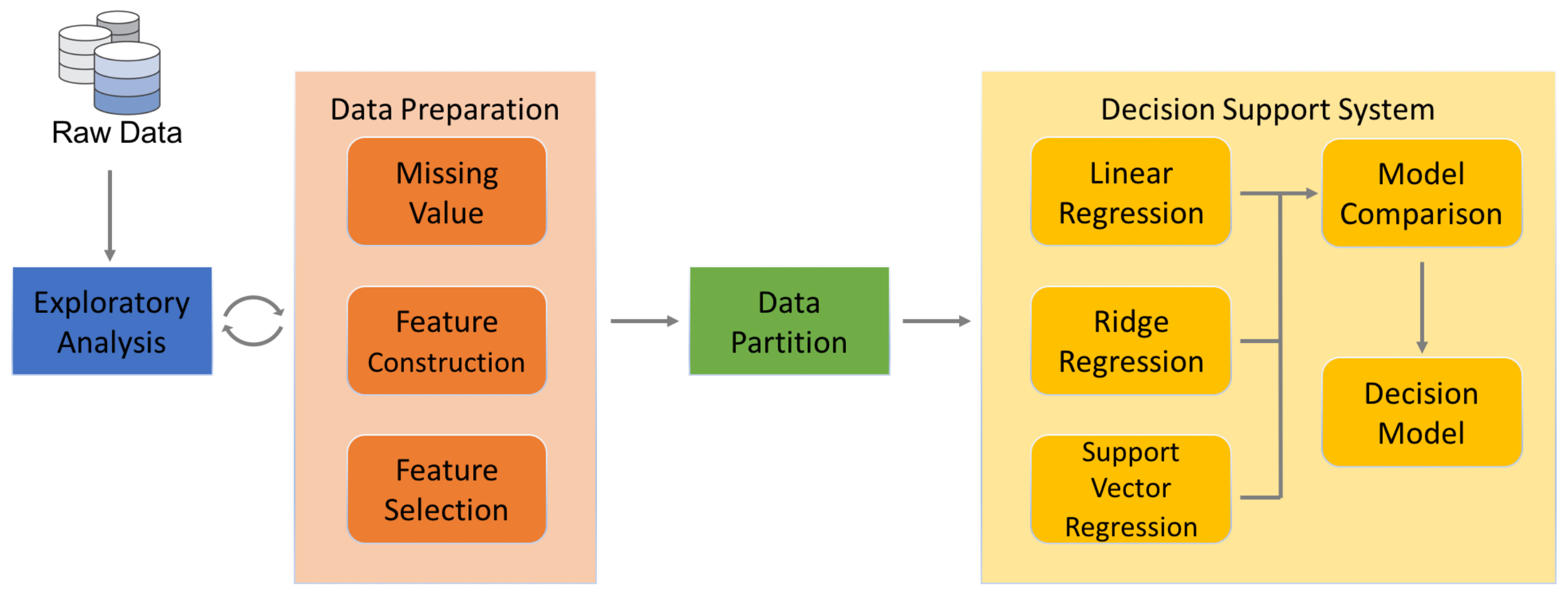


Figure 2: Flow Chart

Exploratory Data Analysis

The project started analysis with obtaining a good understanding of the data. EDA was performed by investigating descriptive statistics and graphical summaries. During this process, it is found that the original raw data was not that helpful. After processing the data and performing correlation analysis, more meaningful features are created to be used in the predictive models.

Data Preparation

To prepare the data for prediction, the project removed the few records where missing values existed, constructed features, and performed feature selection. For feature construction, the project obtained the average total empty vessels' capacity by calculating the empty vessels' capacity from 27 to 21 days ago. After correlation analysis, the feature average_total_empty_tonnage, which measures the average total empty vessels' capacity was the only feature used to predict LPG demand.

Data Partition

As shown in Figure 3, the entire dataset was randomly partitioned into two groups: 80% training data and 20% testing data. The training data was further split into 5 folds (16% in each fold). The model was trained on f1~f4 and validated on f5, and this process was repeated 5 times among f1~f5. The testing data was used to evaluate and compare final performances of models.

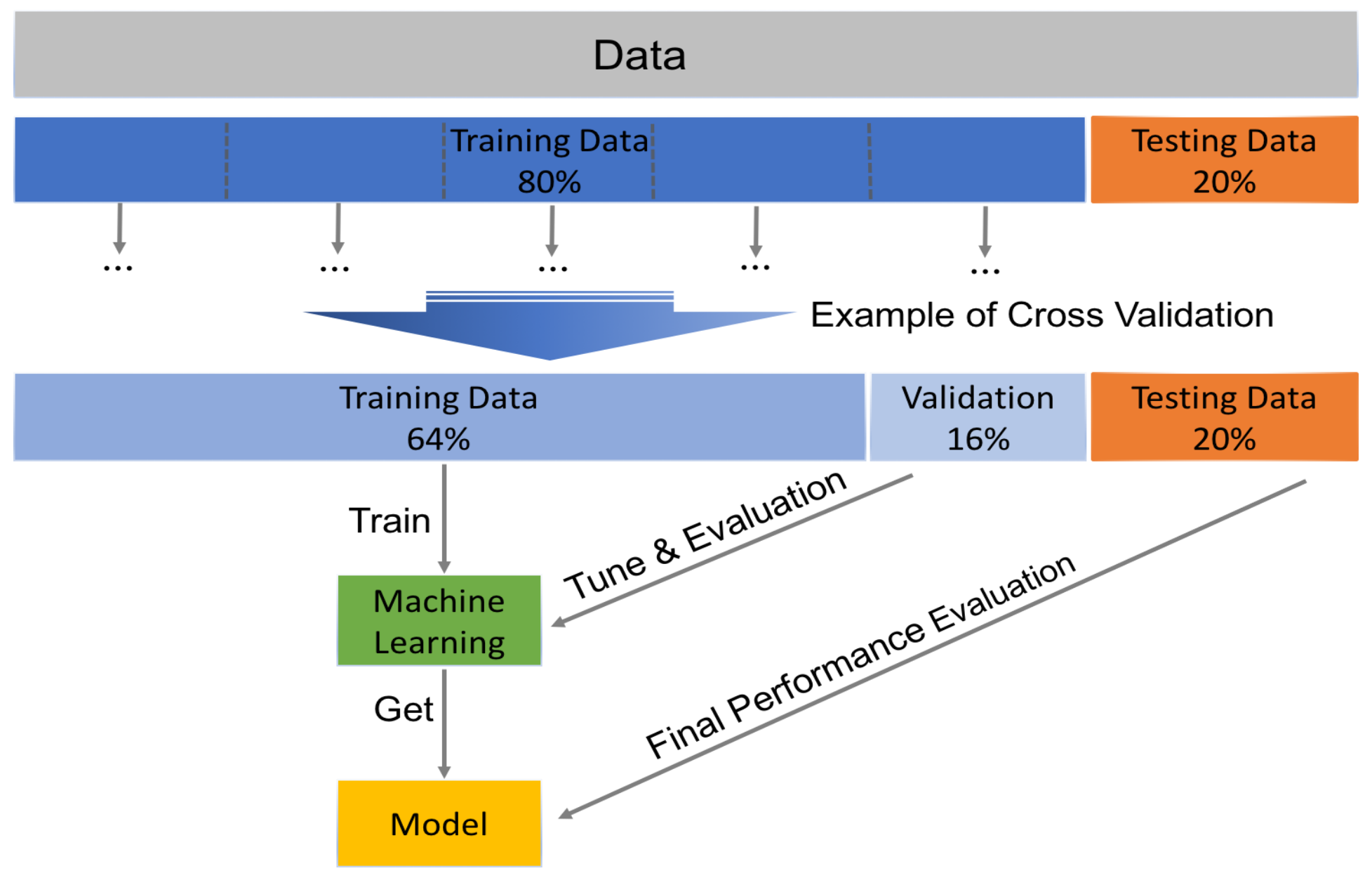


Figure 3: Data Partition

Model Building and Comparison/Selection

To estimate the LPG demand, predictive models were built on the training data using machine learning techniques, namely, Linear Regression, Ridge Regression and Support Vector Regression (SVR). In order to measure the performance of the models, Mean Absolute Error (rMAE) is calculated on the testing data as the most important metric. The smaller the value is, the better the regression model is with predicting future LPG demand.

Decision Model

Based on the metrics selected, R square and the relative Mean Absolute Error (rMAE), it is found that Ridge Regression had the best performance on both train data and test data among the models experimented with. The decision prediction model is used to predict the demand of LPG.

Results

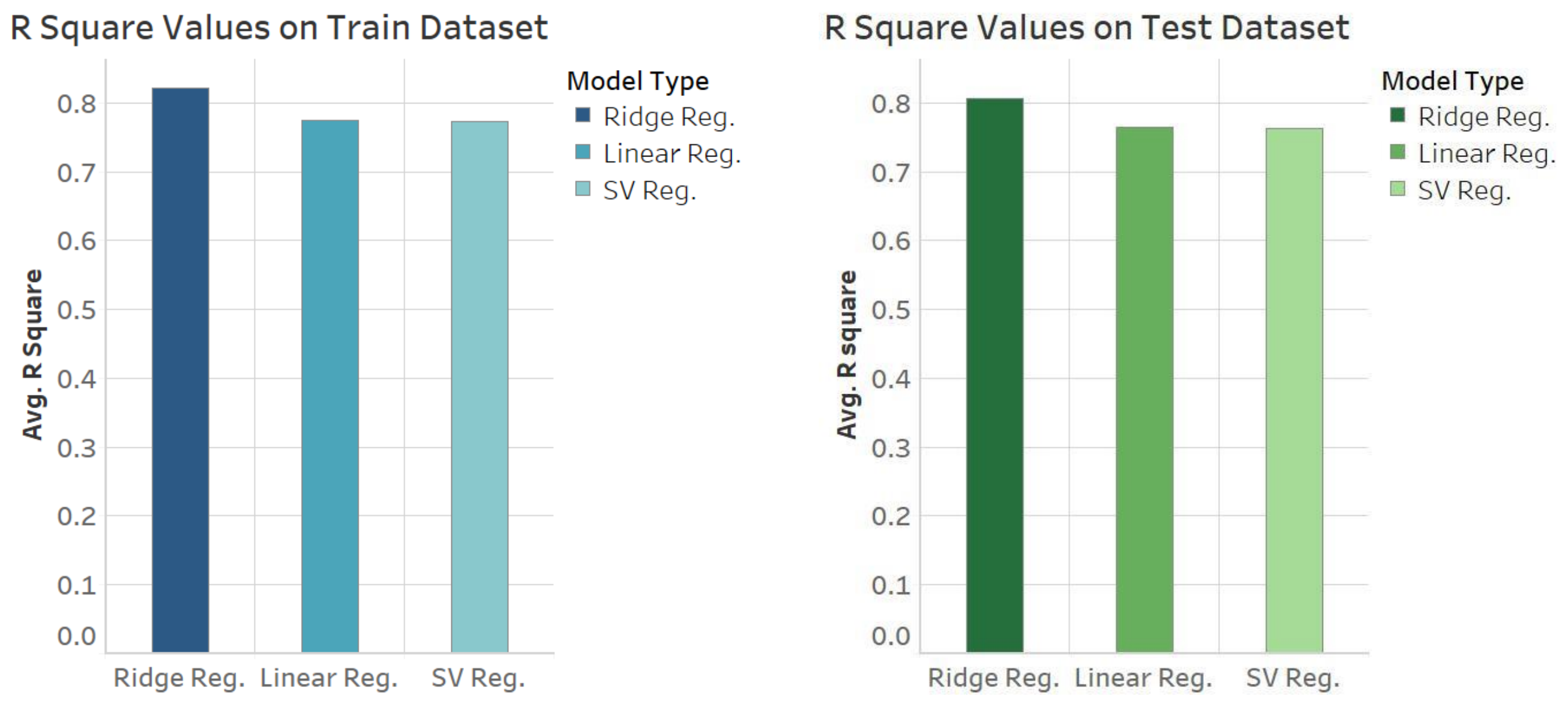


Figure 4: Model Selection Summary

Model comparison

According to the results, two bar graphs were presented in Figure 4. The first bar graph is about the performance of models on training dataset, while the second is on testing dataset. R square is what the project used to measure the performance on both training dataset and testing dataset. The higher R-square is, the better the performance is. To be more meaningful and understandable for the industry, the project used relative Mean Absolute Error (rMAE) to assess the performance of models on testing data. The lower rMAE is, the better the performance is. Based on the rules defined to evaluate the models, it can be concluded that Ridge Regression has the best performance on both training data and testing data.

The Formula of Mean Absolute Error: $\text{Error Ratio} = \frac{|y - \text{pred\_y}|}{y}$

y: actual demand of LPG

pred_y: predicted demand of LPG

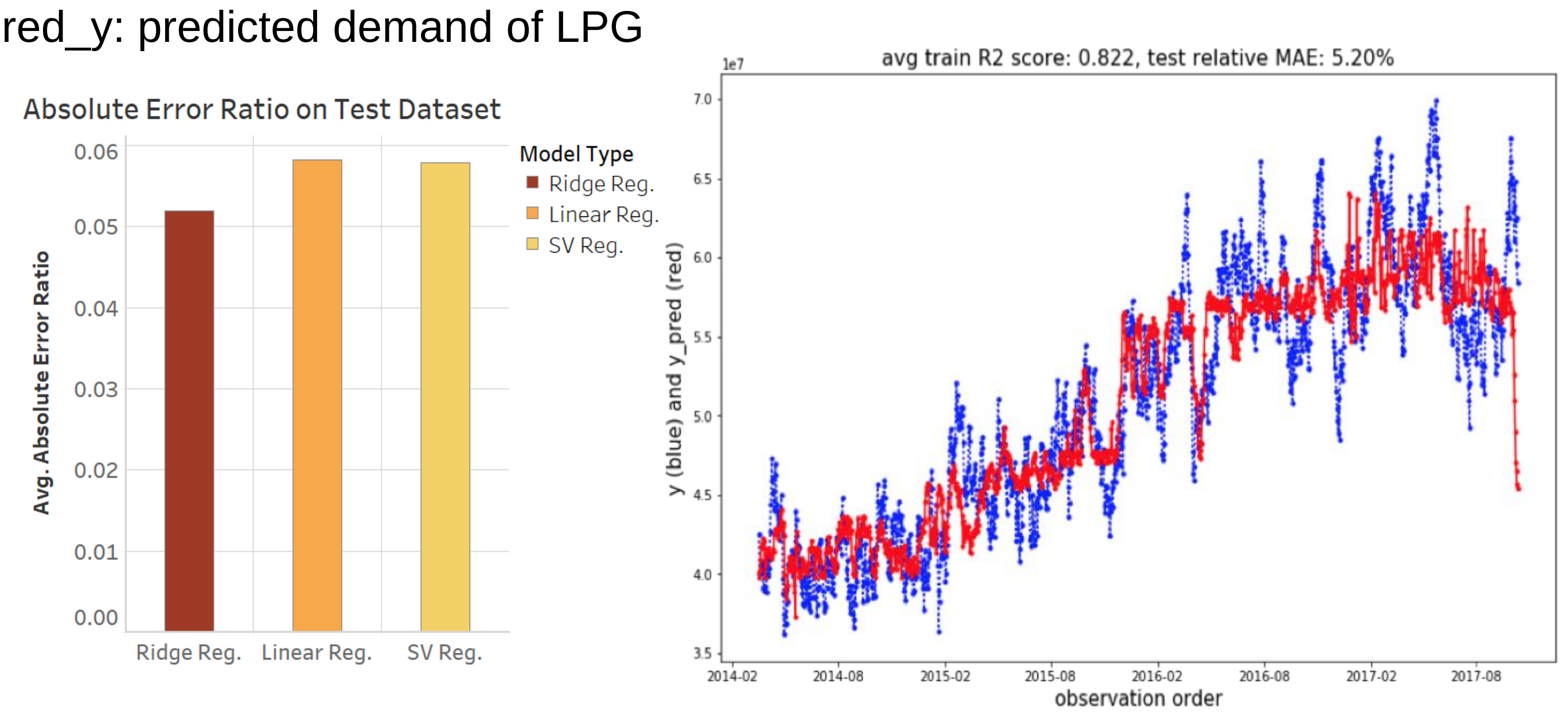


Figure 5: Plot of Actual Values and Predicted Values

Decision Model

In Figure 5, the actual total LPG demand and the predicted LPG demand is shown. The blue line is actual demand while the red one is predicted demand based on Ridge Regression, which has the best performance, two lines having almost the same trend with rMAE of 5.2%.

Figure 6 shows the four-in-one plot for the Ridge Regression model. As shown, all the assumptions of the regression (linearity of residuals, independence of residuals, normal distribution of residuals and equal variance of residuals) have been met, therefore justifying the correct usage of the regression model.

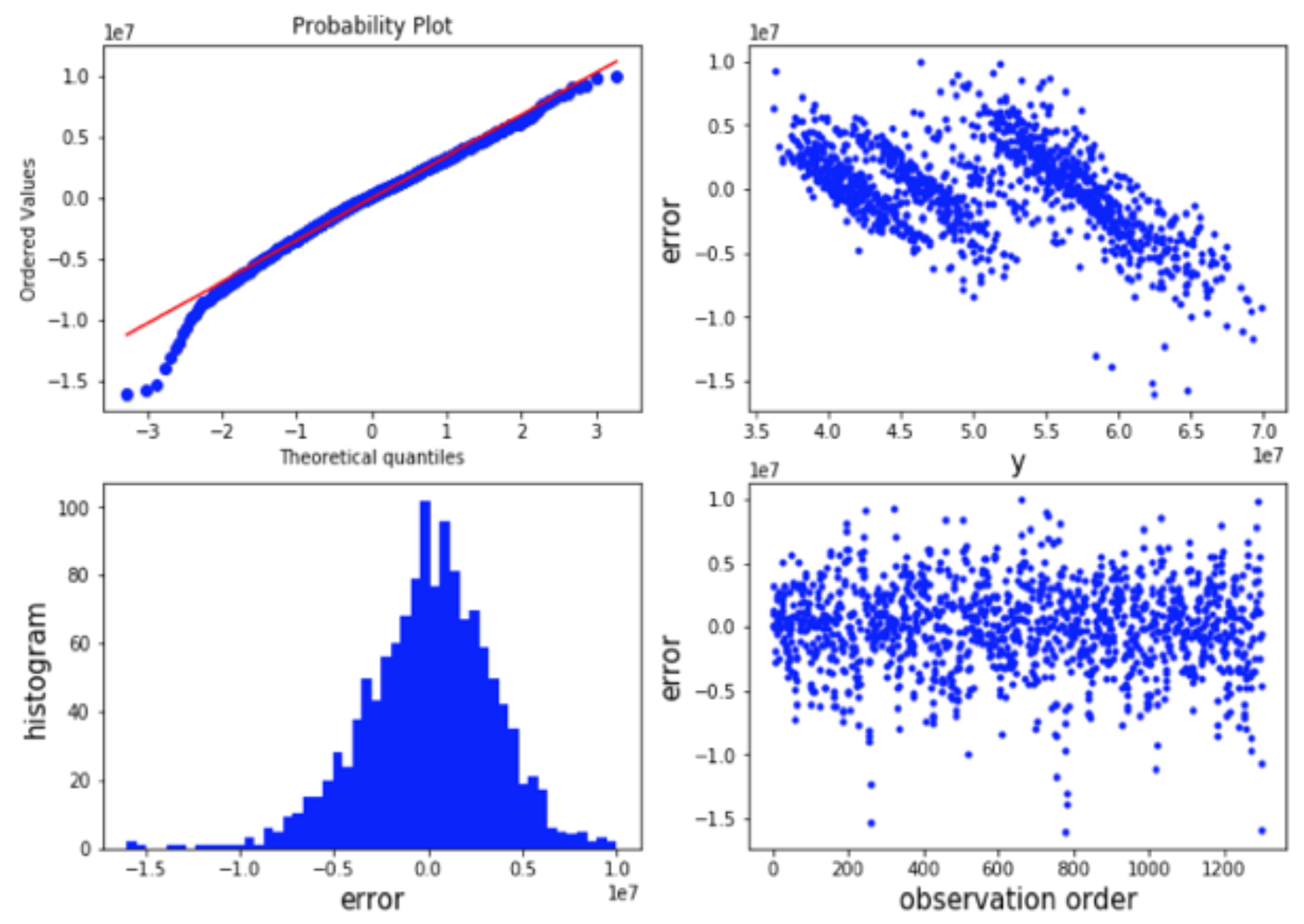


Figure 6: Four-in-one Plot for the Ridge Regression Model

Conclusions

The ability to predict future commodity flows of LPG is very beneficial to companies at both the supply side and the demand side. With better predictions, companies can optimize their operations to either save costs or earn more profit. The model we came up with can help predict the future amount of LPG being shipped three weeks after the detection of empty moving vessels. In the future we can incorporate other features such as seasonality, region, and consumer behavior to help make this model more robust against errors. To sum up, this exploratory step we took in utilizing shipping information has returned valid results; in the future, better refinements will continue to make the predictive model stronger and flexible to industry requirements.

Acknowledgements

We would like to thank our industry partner and professor for the opportunity to work on this project. The Purdue BIAC partially funded this work.