Fuel Price Forecasting

A Mosaic Data Science Case Study

Background

Our client, one of the largest Oil & Gas companies in the world, wanted to improve the process used by industry advisors to forecast fuel price spreads and crack spreads[1] for a variety of products. These forecasts support decisions related to production planning, refinery planning, and open market crude oil trading. The goal was to build automated prediction models to support the forecasting process, allowing advisors to make better informed decisions. Manual, time-consuming processes were predominant. Advisors collected data manually and used spreadsheets to review and obtain insights from the data. This process was repeated several times a month by multiple advisors.

Mosaic, a leading data science consulting firm, was brought to help with the development of the forecasting models and to define the overall architecture to support the production forecast system, including data sources, outputs, and production code for the forecasting models. The predictive models save analysts substantial time on data collection and integration and allow them to make better decisions based on insights provided by the models.

Approach

Mosaic collaborated with customer stakeholders, including industry analysts and IT support staff, to determine the requirements for the models and production system. The models would need to forecast price spreads for a variety of petroleum products up to three months ahead for three different regions in the world.

Mosaic integrated data from multiple data sources. Mosaic designed infrastructure to support predictive models that can process hundreds of variables. These include: supply and demand measures, market prices, crude production forecasts, imports/exports, and refinery outages. Input data is pulled from SAP HANA high-performance in-memory database and flat files, and outputs are stored in SAP HANA and IBM Cognos TM1. Analysts can access TM1 data directly from their systems and easily review forecasted prices. The automated framework facilitates frequent updates of the predictions to boost model performance by using the latest available data.

A variety of advanced machine learning models were evaluated to maximize model performance for each of the forecasted target prices. The production code executes ARIMA, Random Forest, and gradient boosted (GBM) models. The models were trained and carefully cross-validated with over 5 years of

[1] In the Oil & Gas industry, the *crack spread* is the difference between the prices of the crude oil going into the refinery and the finished petroleum products coming out of the refinery. Refinery profitability is largely dependent on the ability of managers to select refinement plans that maximize the crack spread for the quality and type of oil coming into the refinery.

historical data. The models learn from past behavior, and relationships identified in the past together with current values of the variables of interest are used to predict prices for the future months.

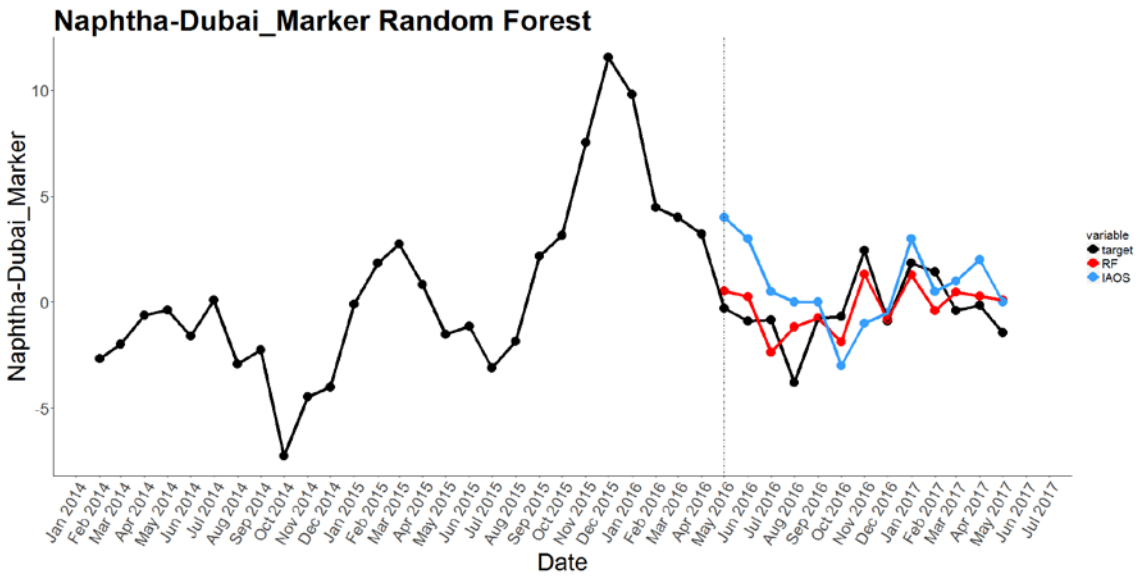

Figure 1 | Performance assessment for a random forest powered forecast showing improved accuracy compared to legacy forecast

The production code was implemented in SPSS Modeler and R, and IBM's Deployment Manager was used to define and schedule productions jobs. In addition to the forecasted values, the production job also generates various performance reports. These facilitate performance tracking and are critical to ensure ongoing model health.

Results

The price prediction models reduced forecasting error more than 50% in most cases compared to legacy forecast methodologies. Industry advisors are now able to make better decisions and increase profits by identifying the most profitable products in the near term and by making better use of the available resources. Additionally, replacing the manual process with an automated solution increased productivity, saving time on data collection and integration, which has allowed industry advisors to focus more time on planning decisions that directly impact the company's bottom line.

For More Information

Want to learn more? Please contact info@mosaicdatascience.com