March 26, 2018

Project Proposal

Sanchit Singhal

Milind Siddhanti

CE 395R – Data Mining with Carlos Caldas

Spring 2018

University of Texas at Austin

Problem Definition

The project will be focused in the petroleum domain and will attempt to come up with patterns and draw inferences between US petroleum statistics about drilling performance and rig counts to the regional oil prices.

Background Research

Recently, the petroleum industry is going through a change not only worldwide but particularly in the US. Although oil prices have fallen over the last decade, America's market share in world oil production has gradually increased. With new found shale oil reserves, the United States is playing a more dominant role in controlling oil prices.

Objective

Given the recent developments, the project will attempt to understand the relationship of the US petroleum production and how it affects the price of oil both internally and externally.

Need

By evaluating the impact of the oil industry on regional oil prices, the US government can better regulate petroleum firms to coordinate a strategy as a country to help them gain a competitive advantage in the global oil markets. Further, companies can utilize this analysis to develop future strategies and direct their investment that maximizes growth opportunities. Lastly, this information can be used to develop economic relationships internally, between states, for a more stable trade market in the US.

Scope

At this stage of the project, it would be premature to define an output but the project team would like to be able to develop relationship between the predictors. Solving a regression/classification problem to predict or forecast either production or price might be a little out of scope - as that can involve thousands of factors - but by drawing inference in the data, the team hopes to learn about the impact of production statistics on regional prices over time.

(Possible) Target Dataset

The dataset consists of features like: Region where the rigs are available, monthly data of the rig counts, production per rig, legacy production change, and total production. The data is available for petroleum liquids, Oil and Natural Gas. Also, features with respect to the rigs such as number of drills that were dug per month, completed and drilled but uncompleted (DUC) are available regionally and these can be used to compare the production rates for completed rigs. Data for the oil prices across the US will be merged to the dataset as an additional attribute and inferences can be made of the captured datasets.

https://www.eia.gov/petroleum/drilling/

Proposed Data Mining approach

At the beginning of the project, the team will begin with data collection by gathering all datasets that they might need. The US Energy Information Administration (EIA) has been identified as a potential data source. Next, the team will go through a selection process in which the target data is finalized. The team recognizes that this may be an iterative process and plans to potentially revisit the data source for more information as and when needed. Once target data has been selected, the team will preprocess the data to ensure it is of high quality and in a format that is easy to use with the chosen software. The transformed data will enable the team to explore the data visually. At this stage, the team will be able to apply various data mining algorithms to extract implicit, previously unknown patterns. Some potential methods include Clustering and Mining Association Rules. By reviewing the results, the team aims to gain knowledge about the relationships.