Sentiment Analysis of GDELT Dataset for Forecasting Crude Oil Price

XIANG LI

GRANT CHARLTON

CS555: DISTRIBUTED SYSTEMS

COLORADO STATE UNIVERSITY

Background Information

- Crude oil price fluctuate due to many reasons
  - Political, military and economic

- Accurately predicting the crude oil price is a challenge
  - Hard to track and analyze the impact of geopolitical events

Problem Characterization

- Global Data on Events, Location, and Tone (GDELT) dataset
  - Contain metadata for all news and articles published worldwide
  - Use CAMEO code to categorize event type

- Big Data & Machine Learning Prediction Trends
  - High school dropouts, Weather, Cyber attacks, Health and disease

Trade-off Space for Solutions

- Dataset Size vs Prediction Accuracy
  - Larger training dataset = more accuracy for prediction model
  - GDELT dataset has terabytes of data

- Realtime vs Offline Computation
  - Realtime computation makes the latest model, but it takes time and affects user's experience
  - Offline computation can serve for more users, but may not keep the model fresh

Methodology - Data

- Using Google BigQuery
- WTI Crude Oil Price

DATE	DCOILWTICO
2012/11/21	87.08
2012/11/22	.
2012/11/23	87.01
2012/11/26	87.28
2012/11/27	86.81
2012/11/28	86.1
2012/11/29	87.64

- GDELT Event

GLOBALEVENTID	SQLDATE	EventCode	GoldsteinScale	AvgTone	NumSources
702903793	20171101	110	-2	2.425755593	1
703000189	20171101	20	3	12.18487395	1
702932720	20171101	23	3.4	-2.09331711	2
703008622	20171101	141	-6.5	-1.69366716	1
703046931	20171101	20	3	-3.84615385	1
703008623	20171101	141	-6.5	-1.69366716	1
702879800	20171101	20	3	2.800819252	3

Methodology – Used Libraries

- Spark Mllib
  - Used for training, testing and evaluating the model
- DataFrames
  - Distributed collection of data = table in relational database
  - Used for storing data from csv files
  - Much efficient than RDD

Methodology - Algorithm

- Data-preprocessing
  - Merge dataset and normalize values by NumSources
  - $AvgToneScore = \Sigma(AvgTone * \frac{NumSources}{TotalSources})$
  - $GoldsteinScaleScore = \Sigma(GoldsteinScale * \frac{NumSources}{TotalSources})$

- Linear Regression
  - data format = (label, [feature1, feature2, …]), dense vector for features
  - $\mathrm{Row}[date] = (price, [AvgToneScore, GoldsteinScaleScore])$

Performance Benchmarks - Setting

- Operating System
  - Linux
  - Program written in Scala
- Dataset Size
  - WTI Crude Oil Price (21.6KB)
  - GDELT Events (5GB)
- Spark
  - worker nodes (10)
  - executor-memory (2G)
  - driver-memory (2G)

Performance Benchmarks - Evaluation

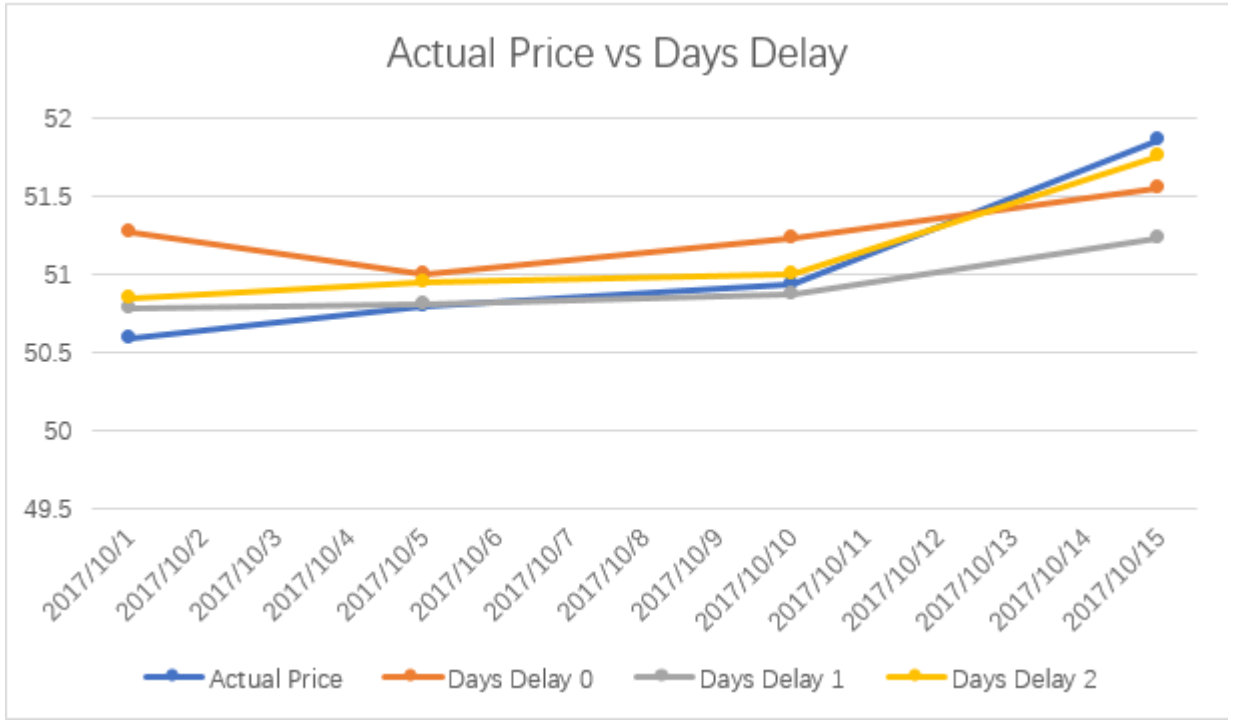
- Actual oil price vs Predicted oil price

- Coefficient of Determination $R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2},$
  - How well a linear regression model fits the data (lies between 0 – 1)
  - It can always be increased by adding more variables into the model

- Root Mean Square Error <= 1

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}.$$

Performance Benchmarks – Days Delay



Key Innovations

- Crude oil price does have a correlation with worldwide events
- Events will affect the crude oil price in 2 days
- GDELT dataset is powerful, it has potential to be used in various ways in the future