
ILLINI DATATHON
ILLINI DATATHON
ILLINI DATATHON
ILLINI DATATHON
ILLINI DATATHON
ILLINI DATATHON
ILLINI STATISTICS CLUB HOSTS
synchrony
DATATHON
ILLINI
February 15-17 2019
Come for prizes, fun workshops, and a chance to network!
Science.
Applied to Life™
LEARN · SOLVE · ANALYZE
Sandia
National
Laboratories
BAYER

265

Next 150 Years

“Provide all Illinois students the opportunity to have a meaningful exposure to data science”

Thank you! Datathon Sponsors:

synchrony FINANCIAL

Sandia National Laboratories

3M

BAYER

PROBLEM STATEMENT

"Predict" the daily stock price for Bayer, Honeywell, 3M and Synchrony from the beginning of 2019 to present using quantitative and qualitative data sources.

Target Variable: daily closing prices

- Quantitative data will be given; includes historical stock prices from 1/1/09 to 12/31/18 or since IPO
- Qualitative data needs to be sourced; may include earnings call transcripts, Twitter API, 8-K, news articles, etc.

Based on your analysis of the data, construct a business thesis on which company or companies you would recommend as an investment and why?

DELIVERABLES

01

predicted values

Daily price predictions Enter closing price predictions on submission_template.csv

02

model

The model developed to predict the target variable

03

business thesis

Based on your analysis, which company should you invest and why? Support with your findings

04

presentation

Top 12 teams will present an overview of their projects. Presentation time should be 10-15 minutes

05

data visualization

Creative or innovative ways of showcasing the data

DATA & RESOURCES

Public Data Sources

- Earnings call transcripts
- SEC Filings
- Google Trends
- Yahoo Finance
- MarketWatch

Python NLP Packages

- NLTK
- spaCey
- Scikit-Learn
- Polyglot
- TextBlob

R NLP Resources

- tokenization
- N-grams
- Topic modeling
- Sentiment Analysis
- Text mining

CRITERIA

All deliverables will be taken into consideration during judging. The following criteria will be on a 5-point scale and will be weighted equally.

ACCURACY

- Based on RMSE
- Takes into account possibility of cheating

METHOD

- Rich model and techniques
- Creative, innovative approach

PRESENTATION

- Engaging presentation
- Clear, thoughtful explanations during Q&A
- Supportive visualizations

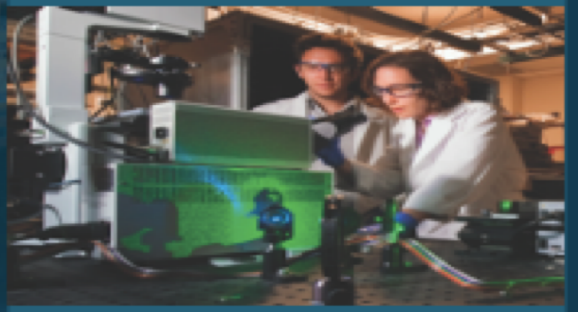
APPLICABILITY

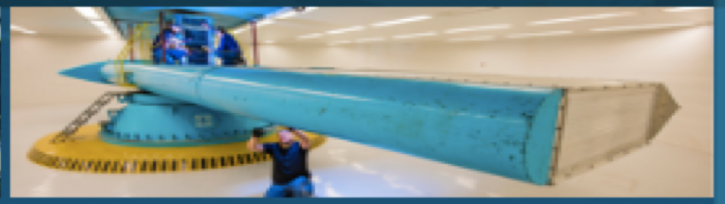
- Solution addresses given business problem
- Proposed solution can be implemented in similar real life situations

RULES

1. Teams are required to develop model(s) based on both quantitative and qualitative dataset (there must be an NLP component)
2. All teams will have 36 hours
3. All teams have until Sunday 8:00 AM to submit all deliverable, early submissions are accepted
4. Submissions will be made through Compass in a clearly labeled zip file stating Team[Number]_datathon.zip
5. Teams can leverage any publicly available data source for generating input features
6. Teams can leverage any tools and/or libraries for this task but are not allowed to ask anyone outside of office hours officials or their own team for help

Principles for Effective Visualization





PRESENTED BY

Andy Wilson

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Principles for Effective Visualization

Concrete Examples: https://venngage.com/blog/chart-design/

- Provide “the greatest number of ideas, in the shortest time, using the least amount of ink, in the smallest space.” (Edward Tufte)
- Show your best estimate plus uncertainty.
- Choose the right chart. Every chart type tells a story.
- Forego decoration. Use color sparingly and for emphasis.
  - Assume your readers are red-green color blind.
- Use the same units between data and chart.
  - Let me make comparisons with my eyes.
- Label the data directly instead of using a legend.
  - Unless the legend is very simple.
  - Or if there are lots of data points.
- Include a complete explanation in text outside the chart.
- Slopes near 45 degrees are easiest to read.