TheDataOpen

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Problem Statement

Welcome to the Fall 2020 West Coast Regional Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background

Over the past few decades, our ability to collect data on consumer preferences has grown exponentially. From music streaming to online shopping, recommender systems have become powerful tools for curating the user experience as well as increasing engagement and revenue. The business value of these systems cannot be understated, with Netflix famously awarding 1 million dollars in 2009 to the team that designed the most accurate recommendation algorithm.

With the rising popularity of streaming in the past decade, the entertainment industry has needed to adapt to the influx of quality TV shows that some call the Second Golden Age of television. As the competition continues to rise and the industry begins to embrace technology, it's important for them to leverage consumer preferences to guide their programming decisions.

Your Task

Your goal is to analyze movie data, potentially in combination with supplementary datasets, to **determine interesting and meaningful implications about consumer preferences**.

We've provided several pre-cleaned datasets relating to movie reviews and associated metadata, but you are not limited in your analysis to these datasets specifically.

You are asked to pose your own question and come up with your own answer using the available datasets in the available time. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight will be rewarded over the breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to investigate your research topic. Submissions may also be illuminating, through the use of data visualizations or through sound statistical tests.

Sample Question 1: Small differences in the design of the systems that collect user ratings and reviews (e.g. MovieLens vs Netflix) have enormous impacts on the data collected. How can we detect these differences and how does it change our interpretation of the data?

Sample Question 2: Are user ratings/reviews predictive of which films/people will be nominated for specific academy awards? How do the ratings change after a movie receives an award?

Sample Question 3: How does variance within a user's ratings tend to affect model accuracy?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and cleaned to be readily usable in most popular data analysis libraries. Note that we will provide you the schema for each of the data tables in another packet.

movie_lense (6)
Ratings from specific users, metadata tags, and genres for movies since 1995.
Size: ~265MB zipped, ~1.2GB unzipped. Source: GroupLens

- **movies**. *58098 rows & 3 columns.*
- **ratings***. ~28 million rows & 4 columns.*
- **genome-tags**. *1128 rows, 2 columns*.
- **tags***. ~1 million rows & 4 columns.*
- **genome-scores***. ~15 million rows & 3 columns.*
- **links***. 58,098 rows & 3 columns.*

the_oscar_award
This table lists all Academy Awards nominations since 1927.
Size: ~900KB unzipped. *10,395 rows & 7 columns.* Source

movie_industry
This table lists popular movies from 1986 to 2016, providing industry metadata about the films.
Size: ~950KB unzipped. *6820 rows & 15 columns.* Source

Additional Datasets

In addition to the provided datasets, you may find the Netflix dataset interesting (TV show data).

You are also welcome to scour the Web for custom datasets to supplement your analysis.

All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One's product team via Slack if you believe your idea is worthy of an exception).

Submissions: Content

Submissions should have three components:

1. Report – this should have two main sections:
   a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
   b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, the use of visualizations is highly encouraged when appropriate.
2. Datafolio – a story-driven visual snapshot of your analysis (please see our guidelines). To start, make your own copy of the template slides provided. **Note that this must be accessible to someone with only a rudimentary understanding of economics, statistics, and machine learning.**
3. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself"**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

Submissions: Evaluation

The competition will have multiple rounds of evaluation. The most important component of this evaluation will be your Report. Of secondary importance is your datafolio. These components are judged as follows:

- **Report Non-Technical Executive Summary**
  - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?
- **Report Technical Exposition**
  - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please

describe your process in detail within your Report.

  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
  - *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

- **Datafolio**
  - *Data Storytelling*: Did you find a compelling narrative that effectively communicates and highlights your analysis? How clear is the phrasing of the question and main insights? Does the layout of the Datafolio complement the flow of the analysis? Please don't just copy and paste text from your report.
  - *Visualizations*: How heavily does the datafolio rely on text when graphics, flowcharts, or other visual representations would have been more effective? How well are the charts and graphs interpreted by accompanying text?

Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

However, please also include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols, please include your LaTeX source file.

Datafolios should also be submitted in **a universally accessible format (PDF, PPT, etc).**

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to upload and send in your submitted content. **Submissions MUST be received by 5:00PM ET on Sunday, September 20th. Any submissions received after that time will NOT be evaluated by the judges**.

Tips & Recommendations

This will be a weeklong event, however, you should try to complete as much of your work as possible before the weekend. The extra time may lull you into a false sense of security. Additionally, with your extra time, you should really think about what problem you want to solve. The outcome of this datathon for you will likely be decided by how well you planned your work.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal + text editor" environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly.

Tips for Success	Try to Avoid
1. Focus on hypothesis testing when brainstorming your research question	**1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy
2. Spend at least 3 hours on your report to ensure strong communication through visualizations and writing	**2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient
3. Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality	**3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile

<u>Ask for Help</u>

Correlation One's technical product team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.