Fall 2023 – INSY 662

Team Project Guidelines and Potential Ideas

This is an opportunity for you and your teammates to apply the knowledge and skills learned in this course to a real analytics problem.

Project Milestones

- Team formation & potential topic(s) – Sep 15
- Project goals and dataset – Oct 3
- Project progress report – Oct 27
- In Class Project Presentations – Nov 21, 23, 28
- Project PPT decks + Peer evaluations due – Dec 5

Description:

Each team of students will identify a particular analytics problem of their choice, along with a data set that will allow you to explore, analyze, and solve this problem. Your analyses should draw on the knowledge and skills you learned in this course. As you explore, analyze, and solve the problem you have chosen, you should feel free to search online and offline sources for information and data, and read articles and research papers that may give you insights into the problem.

The goal of the project is to give you hands-on experience in executing a real analytics project. You and your teammates will go through the iterative process of idea formation, data collection, data cleaning, data exploration, model specification and selection, model estimation, result interpretation, and presentation.

Deliverables:

Each team will deliver its findings via an in-class presentation and also a PPT deck. The presentation will be in class and will run 20 minutes with an additional 10 minutes for questions and answers (may change depending on the number of teams).

Your PPT deck will serve as a written report and should include the following information:

1. Executive summary of your findings and conclusions;
2. Description of the analytics problem;
3. Description of the data source, pre-processing steps, and exploratory results;
4. Description of the model you have selected (*the choice of your model should be restricted to the one covered in class*);
5. Description of the results, how the results should be interpreted
6. Implications of your results and suggestions for businesses, competitors, and markets;
7. References, including data sources, relevant articles and papers you have read.

Evaluation Criteria:

Your presentation and PPT deck will be evaluated based on the following criteria:

1) Benefits that relevant stakeholders can attain from your analytics project
2) Relevance of preliminary steps (e.g., are appropriate pre-processing steps taken?)
3) Robustness of the model (e.g., is the selected model appropriate for the dataset and specified analytics problem?)
4) Ability to translate the results to actionable insights
5) Practicability of your suggestions to the stakeholders (e.g., how feasible is it to implement your suggestions/solutions)
6) Support/justification for arguments you make
7) Ability to communicate your insights and clarity of the report

Topics:
Below are suggestions that make interesting topics for an analytics project. Teams should not be limited by the topics listed below. Teams are encouraged to come up with innovative and interesting topics.

A key factor that limits the selection of topics is the availability of data. Students can ask their previous, current, and future employers, as well as their family and friends, for data.

It is recommended that student teams choose a narrowly focused topic instead of a broad and loosely defined topic. You may also want to limit the amount of data you gather for your project, which may reduce the amount of time it takes to prepare the data for analyses and run your analyses.

POTENTIAL TOPICS INCLUDE:

0. Data from your prior organizations. If you have work experience, your prior organizations may have some relevant problems that could lend themselves to analytics projects.

1. Using Google search to make predictions

Recently, researchers have discovered that consumers' search on Google can be very good indicators of their interest in buying real estate properties. As a result, Google search trends can be powerful predictors of real estate prices.
(https://www.nber.org/system/files/chapters/c12994/c12994.pdf)

Researchers have also discovered that consumers' search on Google can be very good indicators of the attention investors pay to various stocks. As a result, Google search trends can be powerful predictors of stock prices. (http://www3.nd.edu/~zda/Google.pdf)

Can you extend these results to a more fine-grained level (such as for a particular state, city, or industry)? Can you use Google search to make other predictions (such as sales, prices, or incidences of certain events)?

Data sources:

Google search trends (https://trends.google.com/home)

Case-Shiller real estate price index (https://fred.stlouisfed.org/series/SPCS10RSA)

Stock prices (https://datahub.io/collections/stock-market-data)

2. Predict flight delays or rise and fall of airfares

There are publicly available data in the transportation industry. Researchers have identified some factors that can contribute to flight delays.
(https://www.aeaweb.org/articles?id=10.1257/000282803769206269)

Can you extend these results to identify some novel factors that are understudied? Can you extend these results to a more fine-grained level?

Data sources: US DOT airline on time performance data
http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

3. Use border crossing and entry data to make predictions

There are publicly available data on border crossing and entry. Can you use border crossing and entry data to predict the economic activities (e.g., GDP growth, price, trade volume) in countries (e.g., the U.S., Canada, or, Mexico) or various U.S. states?

Data sources:

Border crossing and entry data (https://www.bts.gov/browse-statistical-products-and-data/border-crossing-data/border-crossingentry-data)

4. Predict movie sales

There are publicly available data on movie box office sales. Researchers have studied a number of factors that can drive movie box office sales, including advertising, online movie reviews, timing of releases, etc.

Can you extend these results to identify some novel factors that are understudied (such as the effect of a particular studio/actor/producer, the competition from similar movies)? Can you extend these results to a more fine-grained level (such as for a particular movie genre, summer/holiday season)?

Data sources:
IMDB data (https://developer.imdb.com/non-commercial-datasets/)
Box office data (http://boxofficemojo.com)

5. Use social media data to make predictions

There are publicly available data on social media. Researchers have studied how social media can be used to promote product sales and how social media can predict stock prices. (https://academic.oup.com/rfs/article-abstract/27/5/1367/1581938; https://pubsonline.informs.org/doi/abs/10.1287/isre.2015.0582?casa_token=Q9fIDmbEIp4AAAAA:jdLSlrOiuH73SBxyeLttyriWn7j8AQovX0q7cw8plvi7mhOa4nz5KwGDTX3uL-7oQa76_8E_V04)

Can you apply this to a particular context for which you can get data (e.g., your previous, current, and future employer)? Can you use social media to predict the sales of newly released products or customer satisfaction measures?

6. American Economic Review (AER)

Another great source is the American Economic Review (AER) which is one of the premier academic journals. You can access the journal's paper thru the McGill library. Alternatively, you may access the journal, including past issues, at https://www.aeaweb.org/issues/503. Since all papers published in AER are required to share their datasets, you can find a study that interests you and build upon it. To access the data sets, click on the article title and scroll down to the "Additional Materials" section.

As an example, check this paper which discusses gym membership: https://www.aeaweb.org/articles?id=10.1257/aer.96.3.694. This page contains a link to download the paper, as well as a link through which you can download the data set.

- -

Other Potential Data Sources:

https://www.data.gov/

https://open.canada.ca/en/using-open-data

Economic activity data (e.g., GDP growth, price, trade volume)
http://www.bea.gov

https://www.statcan.gc.ca/en/start

UCI Machine learning repository
http://archive.ics.uci.edu/ml/datasets.html

Kaggle (shows many analytics competitions with actual data sets)
https://www.kaggle.com/datasets

Google Trends
https://www.google.com/finance/domestic_trends

Amazon Public Data Sets
https://registry.opendata.aws/

Movie rating data
http://grouplens.org/datasets/movielens/

NBA statistics
http://www.cs.cmu.edu/~awm/10701/project/databasebasketball2.0.zip

Formula 1 Datasets

https://github.com/toUpperCase78/formula1-datasets

Education statistics

https://data.oecd.org/education.htm

http://nces.ed.gov/datatools/index.asp?DataToolSectionID=4

PewResearch
http://www.pewsocialtrends.org/category/datasets/

Harvard Dataverse (a repository for research data)
https://dataverse.harvard.edu/

Airbnb listings and reviews data
http://insideairbnb.com/get-the-data.html

Instacart order data
https://www.instacart.com/datasets/grocery-shopping-2017

Wharton research data services (need to register)
https://wrds-www.wharton.upenn.edu/

Yelp online reviews dataset
https://www.yelp.com/dataset

Spotify Dataset

https://research.atspotify.com
/datasets/

EU Open Data Portal
https://data.europa.eu/euodp/en/data/

World Bank Open Data
https://data.worldbank.org/