**LINKEDIN DATA SCIENCE PROJECT**

Client: Karl Sjogren, Ad Measurement Lead, LinkedIn

Project Abstract:

LinkedIn is a global employment-oriented business that builds and supports professional networks and connects jobs seekers with employment opportunities.  As part of its portfolio, LinkedIn supports customers in identifying approaches to maximize their connectivity with potential employees and products through analysis of traffic, posts, etc.  However, the large variation in the types of companies, products, customers, and tracked data metrics means that there is no one-size-fits-all statistical approach for tackling these challenges.

For this project, students will gather inputs on common customer requests, statistical approaches, and data problems and build a taxonomy of customer profile segments, related data profiles (eg - periodicity, type, category, volume, and breadth of data set available), known data set (marketing, sales, product).  This taxonomy will be used to help suggest viable, appropriate statistical methods per problem and data profile with validation checklists and document use cases for each method with examples based on provided anonymized data.  This information could be used in an interactive application, where based on user-selected inputs, end-users could input their relevant profile variables for a set of recommended methods, checklists, and high-level procedural documentation.

**NPD PROJECT**

Client: Andrew Dombrowski, Director, Data Science, NPD Group

Project Abstract:

The NPD Group, one of the world’s largest research groups, provides market research that measures and characterizes how consumers shop across all channel types, including both online and brick-and-mortar using a combination of point-of-sale data, tracking, sales, and annual surveys covering the habits of twelve million customers. This consumer data is integrated and parsed from millions of receipts into 36 general merchandise industries from over 50,000 retailers.

Because the incoming data sets are so large, statistical detection methods are needed to help identify when there are data quality issues such as data corruptions, missing values, unexpected changes in data structure or values (e.g. large increases or decreases in sales, amounts, etc). This project will explore using unsupervised learning methods to help identify common data collection errors to help guide further analyst review. Data sources include two years of parsed receipt data from an e-commerce source and an example log of known issues that should be flagged by an automated process.

**NHL PITTSBURGH PENGUINS PROJECT**

Client: Sam Ventura, Director of Hockey Research, Pittsburgh Penguins

Project Abstract:

A common goal in analyzing player performance in team sports like hockey, basketball, and baseball is to develop player performance statistics that estimate a player’s marginal impact on game outcomes, accounting for the strength of a player’s teammates and opponents (among other factors). While traditional box score statistics are informative, they have limitations in terms of evaluating individual performance since those statistics can be heavily influenced by a player’s teammates, opponents, etc.

For this project, we will develop regression-adjusted player statistics for a subset of indicative player data for a data on which talent evaluation and assessment is a goal. In these regression models, the players will be predictors, and outcomes will be goals, shots, turnovers, etc. The coefficients will be interpreted as regression-adjusted statistics for players, which estimate a player’s contribution to his team’s goals, shots, turnovers, etc., while accounting for his teammates and opponents. These statistics will give a complete picture of player performance beyond simply estimating the player’s impact on goals. We will also use these regression-adjusted statistics along with player ages to predict future performance of these players. These metrics will help the Penguins create a process to identify player differentiation and performance as an input to talent assessment.

Because of collinearity issues, regularization methods will be used, starting with ridge regression. Ridge regression is a special case of Bayesian regression, where the coefficients have a prior distribution with mean 0 and a standard deviation related to the shrinkage parameter lambda. Since this special case may not be appropriate for some of these statistics, more general Bayesian regression will be used as well to allow for more flexible regularization. Students will study different choices for prior distributions (e.g. with non-zero means, with informative vs. Subjective priors, etc). Previous knowledge of Bayesian regression is not required.

The resulting posterior distributions for the coefficients will enable the Penguins to identify not only the top players in terms of expected value, but also “high-upside” players, or players with long right tails. These players may have a lower expected value than others, but a higher probability of being very good. Results of the models will be presented as visualizations in an interactive Shiny app.

Data will be provided by project sponsors in a convenient, user-friendly format. Data will contain information required to build the regression models.

The data provided is proprietary to the project sponsor. The data, model, results, and any code referencing data structure are considered confidential information and cannot be shared publicly or in any manner inconsistent with the terms of the Confidentiality and Non-Disclosure Agreements. Students will be required to sign a Form of Participation Agreement, outlining the specific rights of the sponsor and the students with regard to intellectual property and confidential information.

**COVID RESEARCH PROJECT**

Client: Seema Lakdawala, Principal Investigator, Lakdawala Lab, University of Pittsburgh

Project Abstract:

The Lakdawala Lab focuses on developing methods to better understand how influenza viruses occur and adapt as well as their viability in different environmental conditions including airborne transmissibility.  One of their primary goals to develop comprehensive surveillance systems to determine pandemic potential for influenza viruses.  Given our recent circumstances, they have now expanded their work to the COVID-19 virus.

Their current project Public Health Interventions aGainst Human-to-Human Transmission of COVID-19 (PHIGHT COVID) is a publicly accessible data analytics platform that combines disparate public data sources into interactive tools and visualizations.  The current work focuses on the relationship between transmission rates and public health interventions, travel quarantine orders, K-12 school closures and child care status changes, and restrictions of state parks and other public areas.

See <https://lakdawala-lab.github.io/PHIGHT_COVID/index.html> for more.   This research project will involve adding more statistical analysis and modeling to the current set of PHIGHT COVID Maps with initial emphasis on better modeling transmission rates and public health interventions.

This work is also in partnership with public health officials in the Midwest.

**SIMON INITIATIVE RESEARCH PROJECT**

Client: Norman Bier, Director – Open Learning Initiative, Carnegie Mellon University

Project Abstract:

The Open Learning Initiative (OLI), a product of Carnegie Mellon University’s Simon Initiative, provides textbook-replacement courseware built upon principles gleaned from decades of research in three CMU’s strengths: cognitive science, computer engineering and human-computer interaction.

OLI provides materials that can be used in face-to-face, online, or hybrid classes. It’s a platform for delivering high-quality materials with the ability to facilitate groundbreaking research — in technology enhanced learning, data science, learning behavior and more.

In this project, the team will perform exploratory data analysis pre- and post-COVID to understand successful patterns in courses, focused on student outcomes and performance by various demographic subsets. Students will leverage previous analysis and external research to investigate closing gaps on outcome improvement, with potential for publication. The project requires identification of instructor use patters pre- and post-COVID, and analysis to determine the relationship between usage and student outcome.

**IFDA PROJECT**

Client: Annika Stensson, Director of Research & Industry Insights, IFDA

Project Abstract:

IFDA is the premier trade association for the foodservice distribution industry, contributing to the growth, development, and success of the foodservice distribution industry through advocacy, events, research, and education. Their members anecdotally report recruiting challenges in developing sufficient pipeline of qualified external candidates to keep up with their workforce demands in warehouse and commercial, interstate truck driving job families.

IFDA endeavors to develop a deeper, qualitative and quantitative understanding of why target candidates apply for their or other positions, what the key influences are in driving candidates to applications, and what potential remediations are possible for attracting additional candidates to member jobs and careers in order to become the premier destination industry of choice for job-seekers.

In this proof-of-concept project, students will engage with IFDA and a subset of IFDA members in a multi-phased approach to gather and analyze relevant information in conjunction with analyzing demographic and compensation data sets, some provided by IFDA and some publicly available from the US government (e.g. Bureau of Labor Statistics) and propose a roadmap and needs assessment for improving recruitment and retention.

This project includes extensive client-facing work during core hours of business, EST. Preference for students in Eastern and adjacent time-zones, for scheduling meetings with IFDA partner executives.

**NBA Research Project (Basketball)**

Client: Kostas Pelechrinis, Associate Professor, School of Computing & Information, University of Pittsburgh

Project Abstract:

A Bayesian approach to adjusted plus-minus

The development of all-in-one player evaluation metrics have been the holy grail for all sports, including basketball. Adjusted +/- was introduced in the early/mid 2000s and it was a regression-based metric, analyzing *basketball stints* (periods of times where no substitutions were observed). One of the problems of the early versions of adjusted +/- was the collinearities (typically teams use pairs, triplets etc. of players together) that made it hard to assign credit among sets of players. To solve this problem regularized methods were introduced. More recently, these methods have been enhanced with tracking data to increase the predictive power of the metric.

One of the challenges with regularized methods is the challenges underlying the estimation of the uncertainty of the calculated metrics for the players. This is rather important when it comes to questions such as “Is player A better than player B?”. In this project, we will focus on a different approach in estimating the adjusted +/- for a player using Bayesian regression.

Bayesian regression allows us to start with a prior distribution for each coefficient (prior belief on the quality of a player in our case) and then through the data update this belief and obtain the posterior distribution. There are various things to explore, including:

1. Use the player’s contract as the prior. A monetary value can be assigned to a win (and correspondingly to a +/- value) and hence, a contract can be translated to an expected +/- performance for a player. The question becomes how one goes about setting the standard deviation of this prior distribution and whether using this as a prior allows for “quicker” convergence during the season.
2. Movement of the posterior, which will allow us to identify over and under paid players based on their performance.
3. In order to obtain more stable estimates, a typical approach is to use multi-year estimations of the adjusted plus/minus of a player. How does using this “contract prior” for a single season, compares to multi-year versions (without the contract prior).

The data needed for this project are publicly available. They are partially ready to use, and partially needed to be collected. In particular,

The various deliverables for this project will be determined more accurately after the questions to be answered are finalized, but they will essentially include answers to these questions and a blueprint on how to use these results. A website with interactive visualizations – and possibly projections for players value – would be most useful. For example, the website will be presenting the posterior distribution for each player’s +/-, and will be able to probabilistically compare two players based on these distributions.