**Project 1 – Whitehat Jr**

Client: Matt Rogers, Director of Brand Marketing, Whitehat Jr.

Project Abstract:

WhiteHat Jr. is one of the fastest-growing education technology companies in both India and the United States. Their flagship product is a 1-on-1 coding platform that teaches kids from 6-18years-old how to code, regardless of skill level. Offering customized curricula and qualified teachers they have grown to over 30,000 students in the US and perform over 200,000 classes per week worldwide.

While there are general studies that tout the effect that coding can have on a child’s logic, structure, creative thinking, sequencing, and algorithmic thinking, no detailed assessment has been done for their particular platform, curriculum, and teachers. One of their long-term goals is to characterize any impact that their classes have on their U.S.-based students’ lives, including academic, behavioral, and personal. There is anecdotal evidence that the platform is beneficial for students with special needs and might be associated with improvements in children’s concentration and problem-solving abilities.

This project will focus on exploring the data collected thus far on students including interactions with the platform (lessons, pace, etc), performance, and open-ended text comments from parents to look for potential patterns in use that might be associated with positive outcomes. This information can then be used to guide development and design of subsequent surveys to support more robust quantitative analyses in the near future.

.

**Project 2 - NPD GROUP 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.

**Project 3 - NHL PITTSBURGH PENGUINS PROJECT**

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

Project Abstract:

This project will examine the impact on career path, determined by the leagues participated in as part of a professional’s development. There are multiple traditional paths to the professional strata, from Junior League -> NHL Minor League -> NHL, to US High School -> College -> NHL to International-> KHL-> NHL, and other defined paths to the NHL.

The goals of this project is to use the path players take as information and project performance. The understanding in the scouting community is that path matters.

This is anecdotal – the project is intended to establish grounding on this thought.

In this project, teams will use tree models to calculate the probability a given player will go down a path, then model the performance given the path taken, iterating through each year of a career, predicting in a given year the change of playing in league in the development path, while also analyzing specific path taken coming out of the model.

This model will need to presented in a ShinyApp for testing and demonstration purposes.

Methods include Markov chain modeling as a player transitions from one league to the next based on how well they performed currently, while separately modeling performance given the league they end up in. The application should facilitate career simulation based on path chosen (development path).

Player data is available and will need to be scraped from the following public data source:

ELITEPROSPECTS.COM

This is accessible by API. Coding in R is preferred.

**Project 4 - PHIGHT 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.

**Project 5 - 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.

**Project 6 - 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.

**Project 7 - 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.

**Project 8 – PPS – Retention/Mobility Research**

Client: Pittsburgh Public Schools

Project Abstract:

The project examines Promise scholarship use and post-secondary retention and mobility.  What predicts use of Promise scholarships for eligible students?  What factors affect retention at and mobility among post-secondary institutions, for students who use Promise scholarships?  Again, the predictor variables are all internal PPS data, the outcomes (which might be quite diverse) are the NSC.

A successful outcome will help the client understand what in PPS drives post-secondary enrollment, retention, and mobility for students who receive a Promise scholarship.

Input data includes: Pittsburgh Promise Scholarship data, Pittsburgh Promise scholarship used at a post-secondary institution, Demographic information (race, gender, special ed status, English Language Learner status), Enrollment (records of each entrance and exit from a school), Course enrollment and completion (including advanced coursework), Attendance, GPA, Magnet school enrollment, Program participation, Promise scholarship awarded, SAT scores, AP test scores, State accountability test results, Career and Tech. Education (CTE) outcomes (industrial-recognized certificates)

Outcome data you will generate includes: Post-secondary enrollment, retention, and completion (National Student Clearinghouse).

Expect a research-focused project, IMRAD report writing.

**Project 9 – PPS – College/Trade School Indicators Research**

Client: Pittsburgh Public Schools

Project Abstract:

The project examines College and Trade School Enrollment Indicators.  Students will build a prediction model for enrollment in college, trade school, or other post-secondary activity.  Predictor variables would come from internal PPS data; outcome measures from NSC data.

A successful outcome will help the client understand what qualities and experiences in high school will more likely lead to college and trade school enrollment.

Input data includes: Demographic (race, gender, special ed status, English Language Learner status), Enrollment (records of each entrance and exit from a school), Course enrollment and completion (including advanced coursework), Attendance, GPA, Magnet school enrollment, Program participation, Promise scholarship awarded, SAT scores, AP test scores, State accountability test results, Career and Tech. Education (CTE) outcomes (industrial-recognized certificates).

Outcome data you will generate includes: Post-secondary enrollment (National Student Clearinghouse)

Expect a research-focused project, IMRAD report writing.

**Project 10 – HCI – Tutor Research**

Client: Vincent Aleven, Human-Computer Interaction Institute, Carnegie Mellon University

Project Abstract:

The purpose of this project is to investigate whether prerequisite relations among math topics can be detected in standard longitudinal log data from tutors.

Given a math "topic" or skill or unit, what math topics or skills or units do you need to master in order to (smoothly) learn that topic or skill or unit? The main idea would be to treat correlations between performance/learning on some tutor units, or on certain skills, as evidence of prerequisite relations.

The data may not be ideal for answering the question. Students will use data from the standard curricular sequence. There are questions of what metrics to use (large space of possibilities, including AFM-derived measures) and at what level to look for prerequisite relations (units? skills? topics? problem types?) that need to addressed with stakeholders.

The data set used will likely be regarding middle school mathematics, however, the team will be tasked with validating whether or not the approach would generalize. These investigations could become a foundation of "adaptive practice of prior knowledge."

**Project 11 – HCI – Learning Discontinuity Research**

Client: Vincent Aleven, Human-Computer Interaction Institute, Carnegie Mellon University

Project Abstract:

This project is tasked with developing a way to detect learning discontinuities within tutor log data to measure effects of out-of-tutor events. A common scenario is one in which middle school students are using tutoring software in class (in person or online). The teacher "monitors" the class and helps those students who appear to need more help than the software gives them, usually with very short individualized sessions focused on the given tutor problem that the student is working on. The team is tasked with answering how helpful these teacher interventions are to students, do they put students on a different learning trajectory, with respect to the skills that were discussed, and how we might measure effect.

Teams may be to look at the error rate on subsequent opportunities for the skill, or the BKT estimate. Team will validate if there is evidence of faster learning (e.g., in terms of AFM slopes - though perhaps taking into account more information about each opportunity than just right or wrong) following the intervention. Clarifying scope and approach is one project challenge.

Data sets include teacher interactions with students, in the context of learning with tutoring software. In addition to using real data, team will explore this problem with "minimally synthetic" data.

This investigation would have many applications, including measuring effects of student-teacher or student-student (e.g., collaborative learning) or student-parent discussions, for the purpose of giving teachers feedback on their interactions with students (e.g,. on a teacher dashboard), or for research purposes (e.g., is letting students work collaboratively for brief periods of time a useful way of combating student struggle?).

Reading prior research is in scope.