SquarePanda Summer Internship

Data Science at SquarePanda

Tyler Kim

Summer 2021

Contents

1 Company Background

SquarePanda is an EdTech startup that focuses on childhood literacy. Their product includes hardware, in the form of a physical playset with interactive letters, and software, in the form of educational games.

While many people take the ability to read for granted, there is empirical data that shows that a large group of children in America are not able to read proficiently. Specifically, in 2019, the National Assessment of Educational Progress (NAEP) found that 65% of fourth-grade students were not able to read proficiently. Furthermore, 83% of students from low-income families are not able to read proficiently by the third grade. While there have been many modifications made to education curriculum, low reading proficiency across the country shows that the current educational model leaves much to be desired.

Unlike traditional educational systems, SquarePanda relies on research based practices to deliver their content in the most impactful way. Everything from the content, to the ordering of the content, has a very deliberate reason for being. Their curriculum, developed based on the "Science of Reading", along with their hardware and software tools are collectively known as the Square Panda Literacy System.

2 Internship Highlights

2.1 Greenfield USD

2.1.1 Problem Description

My first project, which began the second week of my internship, was an analysis of data from Greenfield Union School District in Bakersfield, California. The data was from a summer enrichment program of elementary school students. I was tasked with creating a presentation of visualizations of the data to be used by the sales team, which would then be presented to administrators of Greenfield USD. This is an important meeting for the company, as a large revenue driver are contracts with school districts that use the Square Panda Literacy System. The summer program that Greenfield USD participated in was essentially a pilot program, and the visualizations that I compiled would be a key component in the sales team's pitch to retain Greenfield USD.

2.1.2 Software Used

Python was the main software tool I used to create my analyses of Greenfield data. I used R to create new data frames that were filtered versions of the original data frame. Finally, I used Google Sheets and the VLOOKUP function to combine columns of multiple data sets on a specific column value.

While I have been working in Python for ≈ 3 years, and I have experience with the pandas and matplotlib libraries, I was able to refine and master my skills with the Greenfield USD presentation. I had used the pandas library before, but it was always in the context of the coursework of a programming class. As such, my prior use of pandas was limited to reading and writing csv's, which meant that my initial implementation was a lot of reinventing the wheel in terms of the data analysis power that pandas possesses. After some prompting from my supervisor, I began to really look into the full capacities of pandas, and I was astonished at how incorrectly I was using the library. Pandas made me abandon a lot of the object oriented solutions that I would have implemented otherwise, and in turn I learned how to mask, aggregate, group, and transform data in pandas.

I also explored the seaborn library for the first time. Seaborn is built on the matplotlib library, and was perfect for the externally facing visualizations I created for the Greenfield data. Seaborn was an immensely useful tool, as its default themes were much more sophisticated and aesthetically pleasing than the basic matplotlib plots. Additionally, it allowed you to perform relatively complex aggregation and summary of a data frame within a function call, which saved me the time of completing those transformations in the dataset. The library also includes a lot of functionality for visualizations of all types, and its a great tool to have in my arsenal for conducting data analysis.

2.1.3 Outcome

While my supervisor and I were very happy with the draft presentation that I had compiled, and thought that there was a lot of powerful visualizations that accurately summarized the game play patterns of the students of Greenfield USD, we ended up having to make major changes in the content of the presentation. While the analyses we put together were a good look into how students used the Square Panda system, the sales team was more interested in an evaluation of students and their growth in literacy. While I had contemplated running those analyses, my supervisor and I had agreed that the small sample size (the summer program was only 1 month long) would make a report on student accuracy growth misleading. However, the competing business objectives of the sales team won out, and we changed the presentation accordingly.

While it was slightly discouraging to see a lot of work go unused, in reflection, it was a good reminder for me to keep the incentives of different teams within an organization in mind. With my data scientist hat on, I had forgotten that the sales team would be using this presentation to try and secure another contract with Greenfield USD. I was creating analyses for a sales team, but doing so from the perspective of a data scientist, which did not create a very compelling story for the sales team to tell. While the outcome wasn't great, I feel like I learned a very valuable lesson on alignment within teams from this project.

2.2 SquareTales Word Logging

2.2.1 Problem Description

The team at Square Panda is largely comprised of people who have pivoted from studying education in academia to working at Square Panda. Their reliance on the Science of Reading and other research on literacy and anatomy are what makes the Square Panda Literacy System so powerful. In that vein, every minute detail of the contents of the curriculum are important to Square Panda.

With that in mind, the second big project that I was tasked with was creating a tabular view of various summaries of SquareTales, one of the forms of content contained in the Square Panda Literacy System. I was given a csv of the words on each page of each book, and was asked to determine information about the books. I was asked to find the amount of times each word occurs, the words that appeared for the first time in each individual book, the total number of occurrences of each word across all books, and the order of unique occurrences of words in each book.

2.2.2 Software Used

This task was perfectly suited for Python. I was able to take advantage of the nested dictionary data structure to map books, page numbers, etc. as keys, and the words or number of occurrences as values. I heavily relied on the `dict.setdefault()` method, which is very powerful for creating a dictionary with repeated keys. I had also never turned a data frame to a csv, so getting experience with that function and its parameters was helpful.

Especially when compared to some of the other tasks I was given, this one felt reminiscent of an assignment I would be given in a foundational programming class. Walking through my code while debugging took me back to the fundamentals of programming.

2.2.3 Outcome

This project seemed to be the one that was the most helpful for the SquarePanda team. Since none of the team members knew Python, they had to parse the text and log each word by hand. This proved to be an immensely time consuming project, which is why this task had been sitting on the back burner for several months.

This project truly showed me the power that comes with knowing a programming language like Python. While I tend to get focused on all of the many things that software helps you build, the benefit that comes from increased efficiency in general workplace tasks goes largely understated.

2.3 Automated Data Analysis

2.3.1 Problem Description

After seeing the visualizations I had put together for the Greenfield USD, the Square Panda client success team had requested that we do a similar analysis for them. However, the context in this case is slightly different. As opposed to a sales team trying to secure a customer, the client success team is a team that is working closely with Clark County School District (CCSD) throughout the course of an entire school year. Their goal with the visualizations that we create is to inform CCSD on their student's usage and accuracy, both at the level of individual classrooms, and the school district as a whole.

2.3.2 Software Used

A lot of this project was re-purposing the old code that I had written for Greenfield USD, and making it work with the data from CCSD. However, I had also built a GUI so that the client success team could accessibly choose the visualization that they wanted to produce, without needing to write any code.

The GUI was built using Tkinter. I've never created a GUI before, but frankly I was shocked at the lack of design customizability that Tkinter offered. The pop-up window looked like something from the early-2000's Internet. If I were to create another GUI in the future, I would probably not use Tkinter.

2.3.3 Outcome

This project is yet to be implemented, as the CCSD school year just began. I will be continuing my consultancy with SquarePanda part time through the fall semester, and will update this section as the project progresses.

2.4 Machine Learning Modeling

2.4.1 Problem Description

In the semester before my internship, I took the primary machine learning course offered at MIT. While I thought that it was a great class, I felt that I did not get enough exposure to the practical implementation of machine learning paradigms. I had some extra time in between tasks, and I had access to thousands of data points of student usage. I decided to try my hand at creating a machine learning model.

I decided that I would train a model to predict a student's accuracy based on their game play make-up. The SquarePanda Literacy System is divided into different sections, which loosely correspond to the skill that they train. I aggregated students usage for each skill area, and their accuracy in games with relevant accuracy tracking. I removed outliers by removing observations that were recorded in less than 5 seconds (accidental taps) or more than 15 minutes (a common data logging error after a student exits the app).

2.4.2 Software Used

I used R and the tidyverse library to manipulate the data the way I wanted it. I used Python to create the machine learning model. Within scikit-learn, a machine learning package built for Python. I used random forest regression as my model type, and used other methods within scikit-learn to compute accuracy and other evaluations of my model's performance.

2.4.3 Outcome

I felt that this project was a really good introduction to the process of creating a machine learning model. I had to research best practices in data cleaning, how to prepare data for entrance in a machine learning model, how to choose a machine learning model based on the problem you are trying to solve and the nature of your data, importance scores, and I learned about different evaluation methods in machine learning models.

The model performed well. On an 80/20 test train split, it achieved an accuracy of .913 on test data, without hyperparameter tuning. Furthermore, on data from an entirely different school district, Dinwiddie USD, the model performed with .934 accuracy. I did not get a chance to test the model on any of the other data that my supervisor had sent me, but hopefully that will be an element of my future work with the company.

3 Reflection

Over the course of my internship, I got very familiar with data analysis and visualization tools. My skills in pandas grew immensely, and I was able to learn to use a lot of other packages too, such as seaborn and scikit-learn. I also became accustomed to working in a start-up, which was a new environment for me. I enjoyed a lot about the culture at a startup, and specifically I felt that interning at a start-up was a great experience because the work I was doing was highly impactful. Oftentimes the work I created was directly presented by me to the team using it, and I got feedback directly. Overall, I thoroughly enjoyed my time at SquarePanda, and am excited to move forward with the team while completing my junior year of school.