Springboard—Data Science Career Track

Capstone 2 Project Ideas

by Tamara Horne

Capstone Project Idea #1

1. **What is the Business Problem?**

   Williamson County Schools (WCS) prediction of kindergarten enrollment and related staffing needs: How can an individual elementary school within WCS predict the number of kindergarteners that will join the school the following year in order to plan for staffing needs ahead of time? What combination of factors would together make good predictors of kindergarten enrollment numbers?

   Currently, the Williamson County school system relies solely on the kindergarten registration process to predict the number of kindergarten teachers they will need for each school year. This requires parents to fill out paperwork and submit it to the school. When parents do not register their children by–say, May of the following school year–staffing decisions are made without complete information. Parents have every right to register their children at any time – including the summer months. The result is that teachers will sometimes be transferred to another school only for the original school to find out in September that they actually need another teacher to replace the one they let go. Conversely, sometimes the school will end up with an extra teacher that they then have to send scrambling for a placement very close to the beginning of a new school year. Neither is ideal. Is there a way the school system could consider additional data – like population, birth rates, percentage of kids in that area that tend to choose private school vs public, etc. to help guide decisions about kindergarten teacher retention from the end of one school year to the beginning of the next? The other grades have the benefit of being guided by the previous year's registration but kindergarten does not.

2. **Who are the intended stakeholders and why is this relevant to them?**

   The stakeholders would be the school district, the principals, the teachers, the parents of kindergarteners, and the kindergarteners themselves. The new prediction method could be used directly by the school district and also by the principal at each school.

3. **Where are the datasets available from?**

   Birth rates by county by year:
   http://www.johnstonsarchive.net/policy/abortion/usac/ab-usac-TN.html
   * would need to determine % born by cut off date, though (Aug 15th or so)
   The useful part of this dataset contains 20 columns(years) of data for each county. I would only use one row: Williamson County.

Under age 5 annually: https://data.census.gov/cedsci/table?g=0500000US47187&d=ACS%201-Year%20Estimates%20Data%20Profiles&tid=ACSDP1Y2018.DP05
Contains data from 2010-2022 but each year is a different dataset. I'd need to pull just the one row of data (population under 5 years old) from each of the datasets.

Data on past kindergarten enrollment numbers for WCS schools - request made, waiting for reply

*Should consider birth rate but also growth of whole county population (people moving here from elsewhere)

**A significant problem I see with this capstone idea would be the birthday cutoff for attending kindergarten. Statistics on births and population are generally stated within a calendar year, but birthday cutoffs are in August.

4. **Initial vision of solution and its evaluation**
    1. **What data science approaches do you anticipate you will use to model the business problem as a data science problem? (Supervised-Regression, Supervised-Classification, Unsupervised, or Hybrid)**

        I anticipate using supervised regression to model this problem - potentially rainforest model.

    2. **How do you anticipate that you will evaluate the performance of each of the data science approaches that you envision?**

        I expect I would try R square, MSE, and MAE to evaluate the performance.

5. **How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?**

    WCS could use this information to make more informed decisions about staffing for kindergarten classes at the end of each school year when preparing for the next. Unnecessary teacher transfers could be avoided, providing a more stable working environment for the teachers and for the students they serve.

Capstone Project Idea #2

1. **What is the Business Problem?**

   Factors influencing educational success: What factors, in addition to class size, influenced the educational success rate of the students involved in the Tennessee's Student Teacher Achievement Ratio (STAR) project?

   The Tennessee STAR project collected data on approximately 7000 students in 79 schools in order to determine whether class size had an effect on educational success. To accomplish this, testing data was collected for the students both in elementary school and then again in high school. Using this data, what other factors can we find to have influenced student success?

   Factors to examine:
   - Gender
   - Race
   - Birth date (one of youngest vs oldest in grade)

2. **Who are the intended stakeholders and why is this relevant to them?**

   The primary stakeholders would be the Tennessee Department of Education and the school districts serving the three counties whose schools were involved in the survey: Nashville-Davidson County, Pickett County, and Fentress County. Knowing whether gender and race played a part in the success rate would give our state government and school districts greater insight into how we are doing in our trek toward equality in those two areas. Knowing whether or not birth rate played a factor would be a benefit to the government when deciding cut off dates for kindergarten enrollment, and also to parents in Tennessee who face the decision of whether to enroll their child in kindergarten the year they become eligible or the year after.

3. **Where are the datasets available from?**

   https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FSIWH9F&version=&q=&fileAccess=&fileTag=&fileSortField=&fileSortOrder=

   Students:
   https://docs.google.com/spreadsheets/d/1q8UG7IWJokQrR2ZlA7KJ1tFmUSjsDcUzUPRv9VTzM8o/edit?usp=sharing
   About 11,000 rows of data

   K-3 Schools:
   https://docs.google.com/spreadsheets/d/1VH6oXSYwHj6x4Lu_yL80rkC8_kYWi6-wvwvR3wMhca4/edit?usp=sharing
   79 rows of data

High Schools:
https://docs.google.com/spreadsheets/d/1WmUBXHEwYyGMK7pnr7Fv36pEdhO-YJ1lWmK_ySaP7Ak/edit?usp=sharing
161 rows of data

Comparison Students:
https://docs.google.com/spreadsheets/d/1SrFieP3BbQlgydLzDfEQ3lMoEoslHSiM6PJYMGQjo1Q/edit?usp=sharing
1780 rows of data

4. **Initial vision of solution and its evaluation**
    1. **What data science approaches do you anticipate you will use to model the business problem as a data science problem? (Supervised-Regression, Supervised-Classification, Unsupervised, or Hybrid)**

        I anticipate using supervised regression to model this problem.

    2. **How do you anticipate that you will evaluate the performance of each of the data science approaches that you envision?**

        I expect I would try R square, MSE, and MAE to evaluate the performance.

5. **How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?**

    I expect the Tennessee Department of Education would use the information to influence future educational policy enrollment decisions, and the school districts would use the information to identify areas where greater advocacy is needed in their particular districts. Parents would use the information to make more informed decisions about the timing of enrolling their individual children in kindergarten.

Capstone Project Idea #3

1. **What is the Business Problem?**

   Public transportation safety data exploration: Gaining insight from the National Transit Database. What trends can we see in the safety of public transportation in the United States? What currently are the safest modes in transportation and what can we predict about the future of safety of various modes of transportation in the future based on the data available? For example, will the current safest mode continue to be the safest or not?

   *The scope will need to be refined and narrowed

2. **Who are the intended stakeholders and why is this relevant to them?**

   Transportation departments both at the national and state levels.

3. **Where are the datasets available from?**

   https://datahub.transportation.gov/Public-Transit/Monthly-Modal-Time-Series/5ti2-5uiv
   65 columns, 133,000 rows, years include 2014-2022

4. **Initial vision of solution and its evaluation**
   1. **What data science approaches do you anticipate you will use to model the business problem as a data science problem? (Supervised-Regression, Supervised-Classification, Unsupervised, or Hybrid)**

      Supervised regression (for predictive analytics)

      Possibly time series forecasting

      Explore possibility of both supervised and/or unsupervised approaches

   2. **How do you anticipate that you will evaluate the performance of each of the data science approaches that you envision?**

      I expect I would try R square, MSE, and MAE to evaluate the performance.

5. **How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?**

   I expect the transportation departments to use this data to make decisions about where to invest money in the future. Should they build more rail or invest in a larger bus fleet? Should they focus on safety improvements for their current fleet or are they already satisfactory?