DTSC 691
Machine Learning
Project Proposal
Haibo Li

Goals of the project

One of the major public high schools' goals is to get students ready for college. In 2019, 73.9% of high school students were enrolled in either four-year or two-year colleges. In Chicago, the percentage is only 63%, lower than national rate. In this study, we will use school progress reports and school profile data published by Chicago Data Portal to study college enrollment rate for Chicago public high schools. By using above datasets, a model will be built to forecast college enrollment rate.

Data description

Data source

Data	Data Source
Chicago Public Schools - School Profile Information SY1617	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/8i6r-et8s
Chicago Public Schools - School Progress Reports SY1617	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY1/cp7s-7gxg
Chicago Public Schools - School Profile Information SY1718	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/w4qj-h7bg

Chicago Public Schools - School Progress Reports SY1718	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY1/wkiz-8iya
Chicago Public Schools - School Profile Information SY1819	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/kh4r-387c
Chicago Public Schools - School Progress Reports SY1819	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY1/dw27-rash
Chicago Public Schools - School Profile Information SY2122	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/2dem-8rq7
Chicago Public Schools - School Progress Reports SY2122	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY2/ngix-dc87
Chicago Public Schools - School Profile Information SY2223	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/9a5f-2r4p
Chicago Public Schools - School Progress Reports SY2223	https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY2/d7as-muwj

There are two types of datasets for this project. School progress report and school profile information. The available years for the datasets are SY16/17, SY17/18, SY18/19, SY21/22, SY22/23. There is no information on School Year 19/20. School Year 20/21 does not have school progress report, so it is excluded from the project. Two types of datasets will be merged on school ID. The datasets also contain elementary and middle schools. Those would be excluded too.

School progress reports has 153 columns and school profile has 91 columns. Columns contains basic information like address and contact information will be dropped. And columns which are irrelative to high schools will be deleted.

School progress report will supply features such as suspension rate, misconduct rate, attendance rate, etc. School profile provides students demography, low-income rate, average ACT score, and college enrollment rate. Initial analysis will be conducted to identify correlations and multicollinearity so that features could be identified.

Software

Software	Prospective Use
Python	Utilizing Python package like numpy, pandas, matplolib scikit learn to conduct data cleaning, exploratory data analysis, Model training and visualization
Jupyter Notebook	Source code editor
Flask	Web application framework to deploy machine learning model
SQLite	Database to store data

Analysis plan & Model Specifications

Analysis description

To build a machine learning model to forecast college enrollment rate, there are several steps to complete this project.

Step 1, data preparation and cleaning. Since there are multiple datasets of each year and two types of datasets, datasets will be concatenated and merged. Columns and rows relating with elementary schools and middle schools will be dropped. Missing data in the dataset will be dealt with. Feature transformation will be conducted for categorical features.

Step 2, Exploratory data analysis will be conducted to understand the data through visualization and statistical tools.

Step 3, Model Training: Identify features, choose training models, optimize hyperparameters and evaluate model performance.

Week 1 goals

- Identify project and proposal drafting
- Obtain data

- Submit project proposal

Week 2 goals

- Data preparation and cleaning
    - Datasets merge and concatenation
    - Deleting irrelevant columns and rows
    - Processing missing data

Week 3 goals

- Exploratory data analysis
    - Statistical summary of the data
    - Data Visualization
    - Conduct value distribution analysis
    - Conduct initial correlation and multicollinearity analysis
- Feature transformation

Week 4 goals

- Machine learning model training
    - Identify features
    - Splitting training and testing data
    - Choosing training model and optimizing hyperparameters
    - Model evaluation

Week 5 goals

- Building web application to deploy machine leaning model

Week 6 goals

- Making a presentation video of less than 30 minutes

Week 7 goals

- Final documents preparation
- Submission of the project

Delivery plan

- All the codes used to build machine learning model in Jupyter Notebook formation

- Python code for building web application of model deployment
- Project walk-through video