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 and Week 3 goals

Merge school profile and progress reports of the same year by school ID. Then concatenate the merged dataset together. Drop columns which has nothing to do with high schools. Convert students’ demographic data into percentage.

In order to understand data better, plot graduation rate and some other features using matplotlib and seaborn.

Examine the missing data for the columns. If missing data percentage is over 25% of the column total, drop the columns. Then I am going to look at the categorical columns and numerical columns individually.

For numerical columns, plot histogram of each column to check outliers and distribution. Drop outliers. For Skewed data, I will try logarithm and square root transformation. Later examine collinearity with heatmap. Drop columns which have high collinearity.

For categorical columns, conduct one-hot encoder to transform the data.

I will deal with missing data. I plan to forecast missing data using KNN algorithm.

Finally I will apply scaler to the numerical data.

## Week 4 goals

First, split data into training and testing data

I plan to explore linear regression, random forest regressor and gradient boosting methods to train the training model. I will use grid search and cross validation to fine tune hyperparameters. MSE will be used for model evaluation.

## Week 5 goals

Using Flask to build a web application to deploy machine learning model. Users could use the web application forecast school’s college enrollment rate by providing certain features’ information.

## Week 6 goals

Making a presentation video of less than 30 minutes

## Week 7 goals

Final documents preparation and 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