PUBLIC SCHOOLS IN ONTARIO

- MUKTA JATHAR

CONTENTS

- Objective
- Overview
- Data Sources
- Descriptive Analytics – Tableau
- Predicting school performance – Machine Learning
- Conclusion & Next Steps

OBJECTIVE

Analyze performance of publicly funded schools in Ontario as a function of characteristics of the school & population

Exploratory analysis (Tableau)

Machine Learning to predict school performance

DATA SOURCES

- **Schools** :
  - **Source** : Ontario Open Data
  - **School characteristics**: Board name, type, school type, language, enrolment, location, etc.
  - **School performance**: Performance of the schools is based on EQAO's (Education Quality and Accountability Office) provincial tests that assess student literacy (reading and writing) and math skills at three points in their kindergarten to Grade 12 education.
  - **Student demographics:** Knowledge of official languages, % students coming from low income households, % students whose parents have some university education, etc.

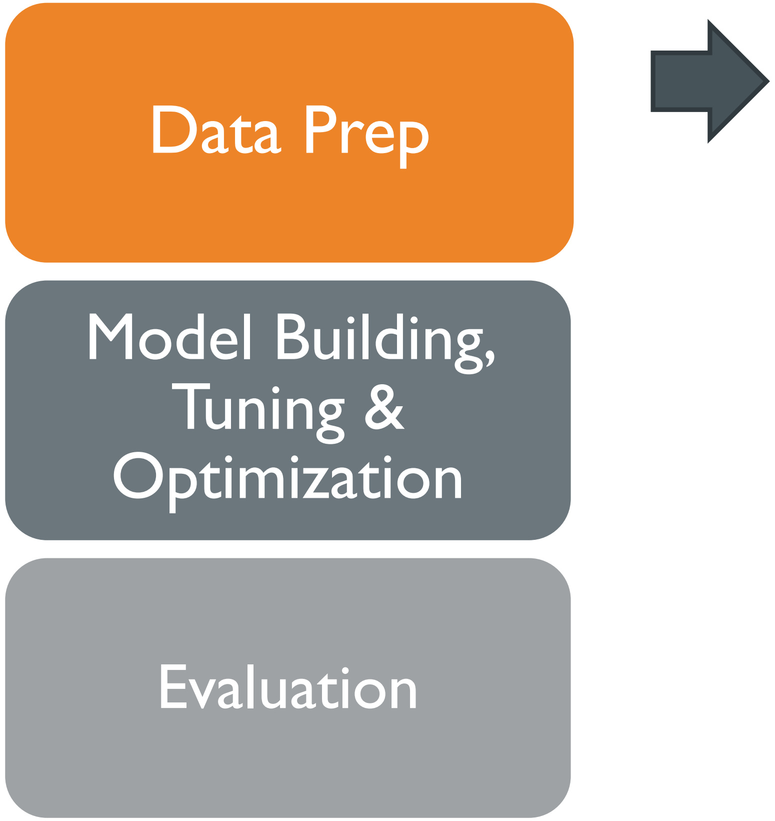
- **Census** :
  - **Source**: 2016 Census (Pulled from CHASS Canadian Census Analyser)
  - **Population characteristics**: Median income, average age, ethnic origin, unemployment, etc.

DESCRIPTIVE ANALYTICS - TABLEAU

Tableau Storyboard

https://public.tableau.com/profile/muk8640#!/vizhome/PublicSchoolsinOntario/PublicSchoolsinOntario

PREDICTING SCHOOL PERFORMANCE – DATA PREP



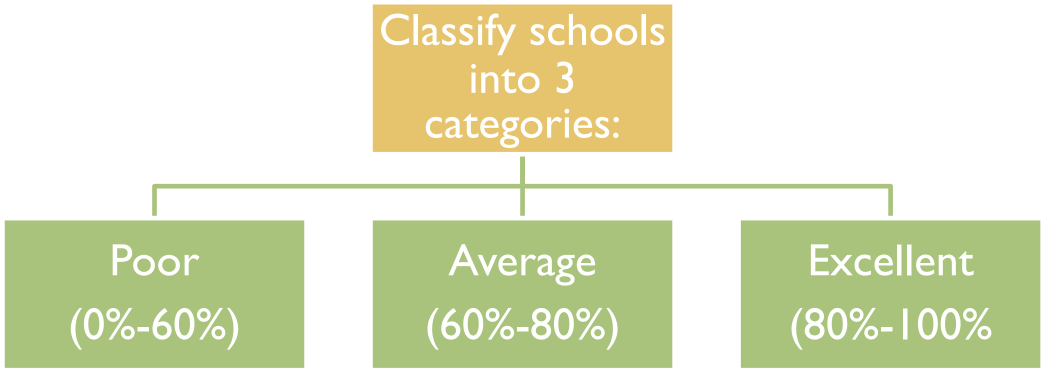
- Data cleaning & preparation :
  - Handling missing values
  - Renaming columns
  - Removing extra columns
  - Calculating average performance schools, etc.
- Mapping census data to school using location
  - Map census areas to the school using nearest latitude and longitude. (Data by school catchment areas not available)
  - Used 'scipy.spatial.KDTree' to find nearest neighbor
  - Calculate the population metrics for the school

PREDICTING SCHOOL PERFORMANCE

Data Prep

Model Building, Tuning & Optimization

Evaluation



Classify schools into 3 categories:

Poor (0%-60%)

Average (60%-80%)

Excellent (80%-100%

Algorithms used:

- Logistic Regression
- XGBoost
- Random Forest

Grid Search for tuning and optimization

PREDICTING SCHOOL PERFORMANCE

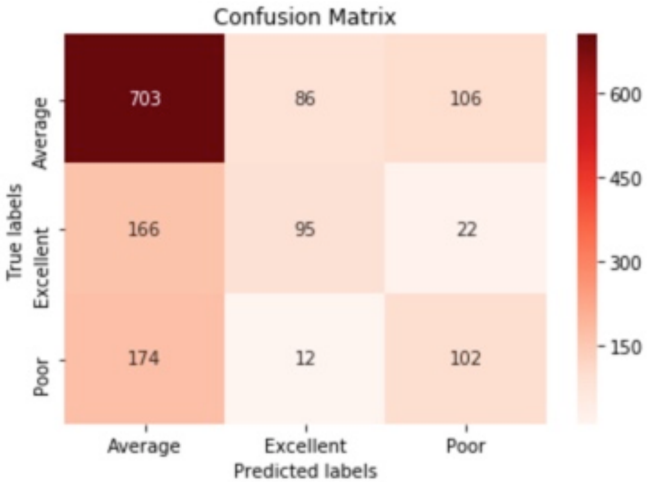
Data Prep

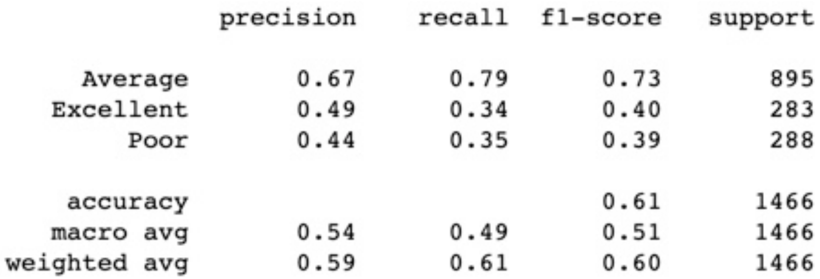
Model Building, Tuning & Optimization

Evaluation

Logistic Regression

Accuracy : **62%**

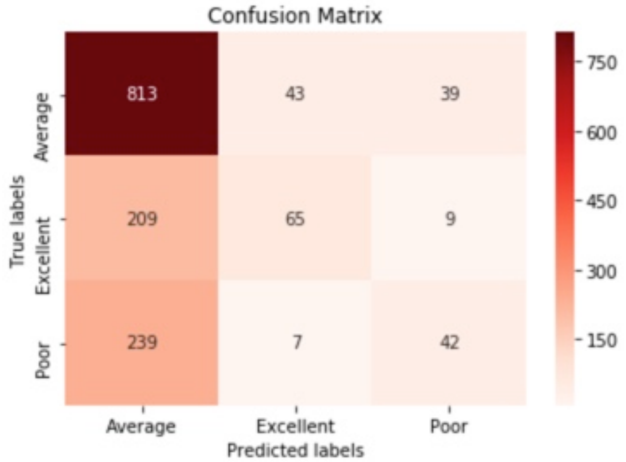


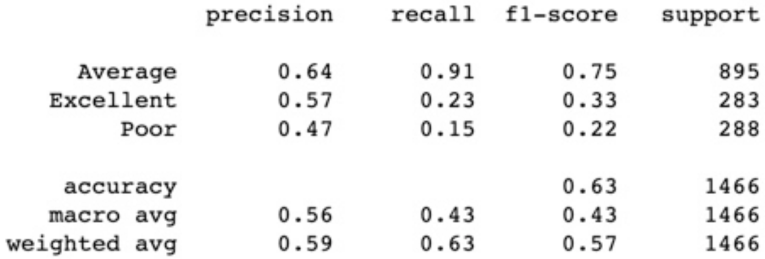


```
              precision    recall  f1-score   support

     Average       0.67      0.79      0.73       895
   Excellent       0.49      0.34      0.40       283
        Poor       0.44      0.35      0.39       288

    accuracy                           0.61      1466
   macro avg       0.54      0.49      0.51      1466
weighted avg       0.59      0.61      0.60      1466
```

XGBoost

Accuracy : **63%**





```
              precision    recall  f1-score   support

     Average       0.64      0.91      0.75       895
   Excellent       0.57      0.23      0.33       283
        Poor       0.47      0.15      0.22       288

    accuracy                           0.63      1466
   macro avg       0.56      0.43      0.43      1466
weighted avg       0.59      0.63      0.57      1466
```

Important Features:

- % Students whose parents have some University Education
- Average change in performance over last 3 years
- Grade Range
- % Students from low-income households

CONCLUSIONS & NEXT STEPS

- Current model tends to misclassify ‘Poor’ and ‘Excellent’ schools as ‘Average’.
- Work on rectifying class imbalance
- Hyper-parameter tuning for optimization
- Revisit features
- Other data points which might help: school budgets, teacher performances, etc.