College Recommender with big data

Name:
Wei Jun Li , Haochen Song, Yuanzhe Liu, Yi Xu, Yuming Xie

Presentation Overview

1 Introduction

2 Literature Review

3 Methodology

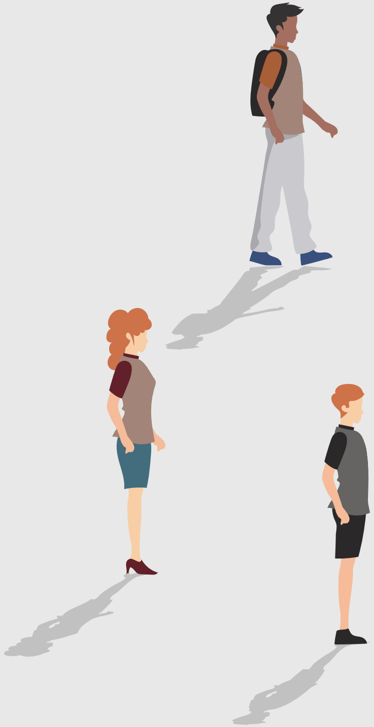
4 Experimental Setup

5 Results

6 Recommender System

7 Discussion

8 References & Appendices



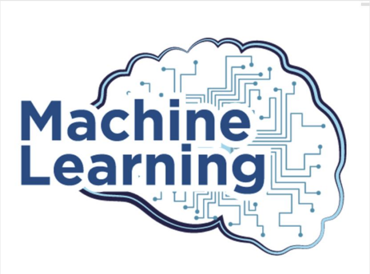
Introduction 01

Background and Motivation

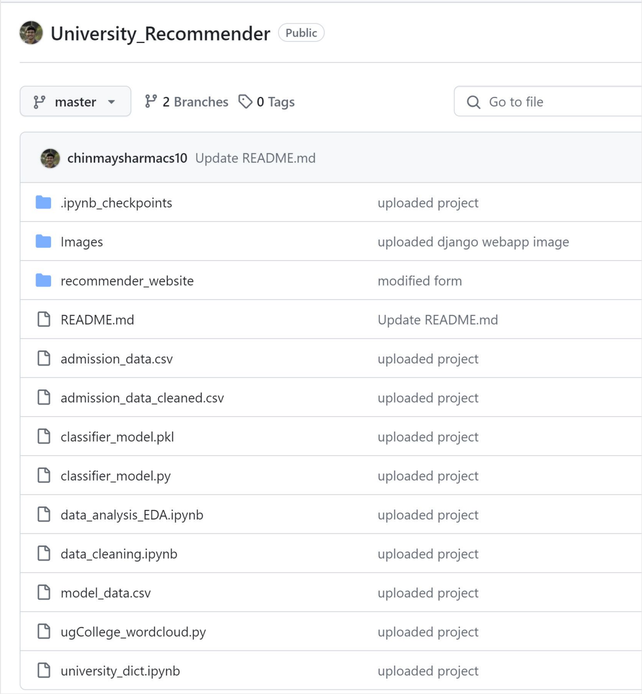
Problem Statement

Objectives and Scope

Background and Motivation



+


University_Recommender Public

master | 2 Branches | 0 Tags | Go to file

chinmaysharmacs10 Update README.md

.ipynb_checkpoints	uploaded project
Images	uploaded django webapp image
recommender_website	modified form
README.md	Update README.md
admission_data.csv	uploaded project
admission_data_cleaned.csv	uploaded project
classifier_model.pkl	uploaded project
classifier_model.py	uploaded project
data_analysis_EDA.ipynb	uploaded project
data_cleaning.ipynb	uploaded project
model_data.csv	uploaded project
ugCollege_wordcloud.py	uploaded project
university_dict.ipynb	uploaded project

36 x

Problem Statement

29 features

Academic performance

Median Earning

Student-to-Faculty Ratio

- ...

<u>Individual needs</u>

<u>Future income</u>

Objectives and Scope

Data Collection and Processing

Sample X 72238

Collected Dataset X 8986

Feature X 29

3 Imputations
- Simple Imputation (Median)
- KNN Imputation
- Iterative Imputation

Model Development and Application

baseline predictive models
- Linear Regression
- Ridge Regression
- Lasso Regression

advanced models
- XGBoost
- MLP
- Random Forest
- Feature Token Transformers

Performance Evaluation and Comparison

3 Imputations & different regression and machine learning models

Analysis and Implementation of Results

predict university fit and potential income

Literature Review 02

- Study 1: "Developing and Evaluating a University Recommender System"
- Study 3: Systematic Review of Recommendation Systems for Course Selection

Utilizes diverse metrics such as Diversity, User Satisfaction, and Novelty to evaluate recommendation quality, reflecting the complex preferences of users in university selection.

- Study 2: "A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data"
- Study 4: A Comprehensive Survey of Recommender Systems Based on Deep Learning:

Focuses on the critical role of data preprocessing and hyperparameter tuning in improving the accuracy of machine learning models, optimizing recommendations for undergraduate programs.

Methodology 03

Datasets

Data clean

Algorithms and Techniques

Tools and Technologies

Justification for the Approach

Datasets

U.S. DEPARTMENT OF EDUCATION
College Scorecard

2012 to 2022 = 10 years data

Sample X 72238

Collected Dataset X 8986

Feature X 29

```
INSTNM               0
PREDDEG              0
SATVR25          60203
SATVR75          60203
SATMT25          60149
SATMT75          60149
ACTCM25          59824
ACTCM75          59824
MD_EARN_WNE_P10  40537
STUFACR           7714
UGDS_WHITE        7598
UGDS_BLACK        7598
UGDS_HISP         7598
UGDS_ASIAN        7598
UGDS_AIAN         7598
UGDS_ASIAN        7598
UGDS_2MOR         7598
UGDS_NRA          7598
UGDS_UNKN         7598
IRPS_WHITE       21651
IRPS_BLACK       21651
IRPS_HISP        21651
IRPS_ASIAN       21651
IRPS_AIAN        21651
IRPS_NHPI        21651
IRPS_2MOR        21651
IRPS_NRA         21651
IRPS_UNKN        21651
PCTPELL           7938
OPEFLAG              0
dtype: int64
```

Data clean

1. Delete null
2. Avoid Data Bias
3. Increase Data Size
4. Fill Other Nulls
5. Output a Cleaned Sub-dataset

Sample X 9197

Cleaned SubDataset X 1455

Feature X 29

```
INSTNM             0
PREDDEG            0
SATVR25            0
SATVR75            0
SATMT25            0
SATMT75            0
ACTCM25            0
ACTCM75            0
MD_EARN_WNE_P10    0
STUFACR            0
UGDS_WHITE         0
UGDS_BLACK         0
UGDS_HISP          0
UGDS_ASIAN         0
UGDS_AIAN          0
UGDS_ASIAN.1       0
UGDS_2MOR          0
UGDS_NRA           0
UGDS_UNKN          0
IRPS_WHITE         0
IRPS_BLACK         0
IRPS_HISP          0
IRPS_ASIAN         0
IRPS_AIAN          0
IRPS_NHPI          0
IRPS_2MOR          0
IRPS_NRA           0
IRPS_UNKN          0
PCTPELL            0
OPEFLAG            0
dtype: int64
```

Algorithms and Techniques

baseline predictive models
Advanced Machine Learning Models
Feature Token Transformer
Clustering and Data Preprocessing

Tools and Technologies

Python
Jupyter Notebooks
Scikit-Learn, TensorFlow, and XGBoost Libraries

Justification for the Approach

Diverse Algorithms
Comprehensive Dataset
Feature Token Transformer

Experimental Setup 04

1. Experimental Design & Objectives:
    - Focused on predicting university suitability and future income using a dataset of 29 features.
    - Aimed to forecast annual incomes based on SAT/ACT scores, race, and college data.
2. Data Preprocessing:
    - Cleaning: Removed records with high nulls in academic performance and racial categories.
    - Encoding: Applied one-hot encoding to categorical variables.
    - Imputation: Tested various techniques like SimpleImputer and KNNImputer for filling missing values.
3. Model Setup & Tuning:
    - Utilized models like linear regression, XGBoost, and MLP.
    - Optimized parameters using cross-validation for balance and accuracy.
4. Clustering & Ensembles:
    - Employed K-means for clustering and adjusted Random Forest settings based on performance metrics.
5. Evaluation & Validation:
    - Used MSE and R^2 to evaluate model performance.
    - Ensured robustness through k-fold cross-validation.
6. Recommender System:
    - Employed ECLAT algorithm, treating each record as a transaction to predict earnings based on academic and institutional data.

Regressions (Benchmark) 04.1

Tasks:
1. College Application Prediction.
2. Post-Graduation Outcome Prediction (ie. Annual Income).

Models & Parameters:
1. Linear Regression.
2. Ridge Regression (alpha = 1.0). [alpha controls the magnitude of the L2 penalty term]
3. Lasso Regression (alpha = 0.1, max_itr = 1000). [alpha controls the magnitude of the L1 penalty term]

Significance:
1. It serves as base model for future exploration and analysis.
2. It offers intuitive guidance to the more robust and complex models.

Regressions (Benchmark) cont.

Result for Task 1: College Prediction

Model	MSE	R^2
Linear	175042.66	0.0485
Ridge	**174723.56**	**0.0502**
Lasso	174868.41	0.0495

Result for Task 2: Outcome Prediction

(more on next slide)

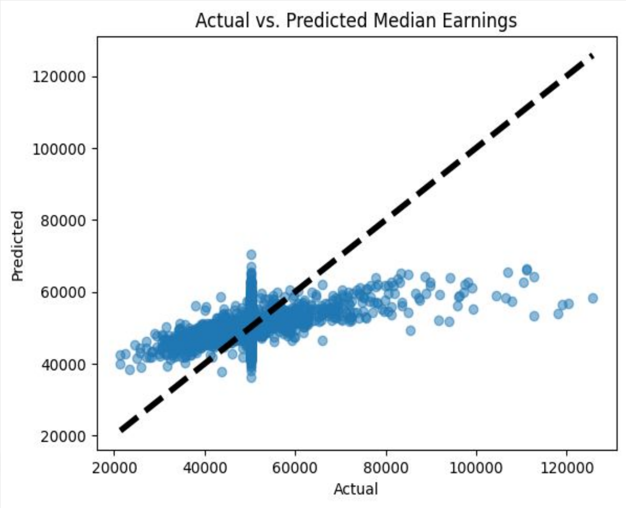
Model	MSE	R^2
Linear	**63243.56**	**0.24**
Ridge	63248.81	0.24
Lasso	63262.79	0.24

Interpretation: The relatively **high MSE** and **low R^2** indicate that the Regression models **may not be** fitting the data very well, we need more powerful and robust models.

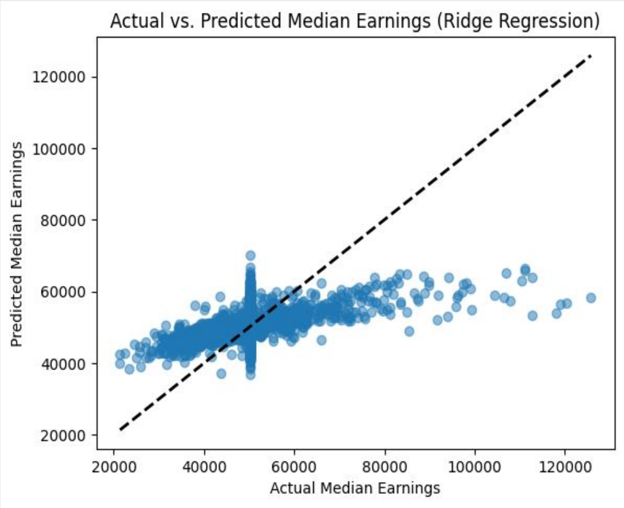
Regressions (Benchmark) cont.

Below are the comprehensive visualizations of the actual versus predicted median earnings using Regressions.

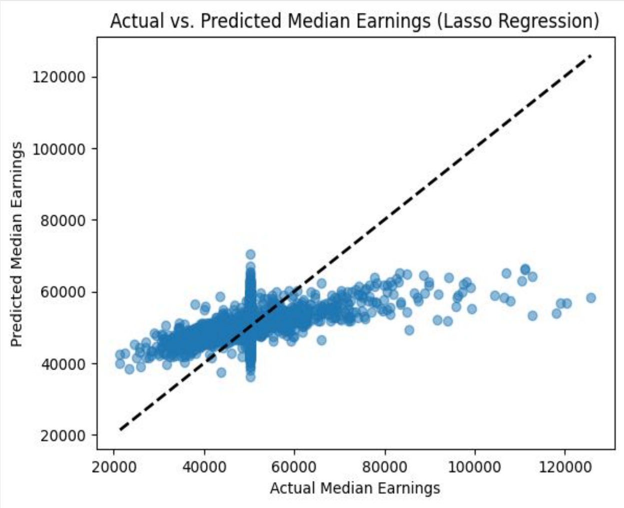
Linear Regression



Ridge Regression



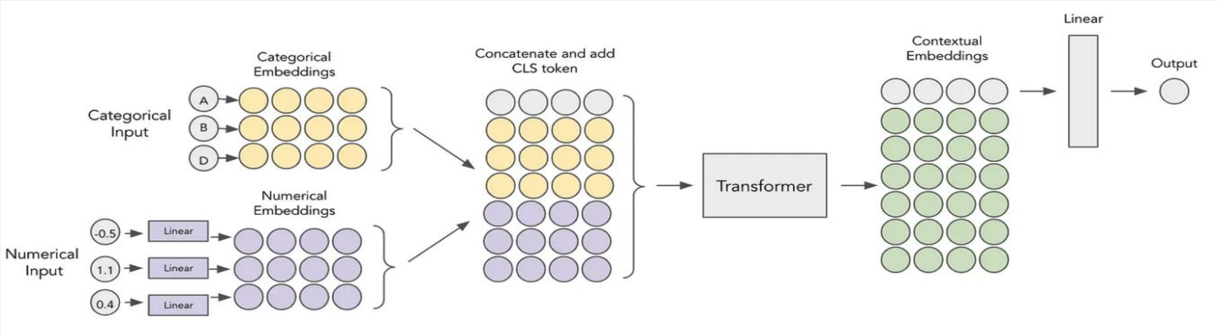
Lasso Regression



Regression Cont: More Model, More Data

Model:

- XGBoost
- MLP (3 layers, Adam with lr 1e-3, 400 epochs)
- Feature Token Transformer



Regression on manually cleaned data

Model	RMSE	R2
XGBoost	8170.32	0.2
MLP	8652.67	-0.89
Feature Token Transformer	7275.65	-0.18

Regression on Simple Imputation

Model	RMSE (MIN)	R2 (MAX)
XGBoost	13728.89	0.03
MLP	13380.84	-45.90
Feature Token Transformer	13577.16	-0.01

Regression on KNN Imputation

Model	RMSE	R2
XGBoost	13649.53	0.55
MLP	19325.81	-16.51
Feature Token Transformer	16153.31	-0.33

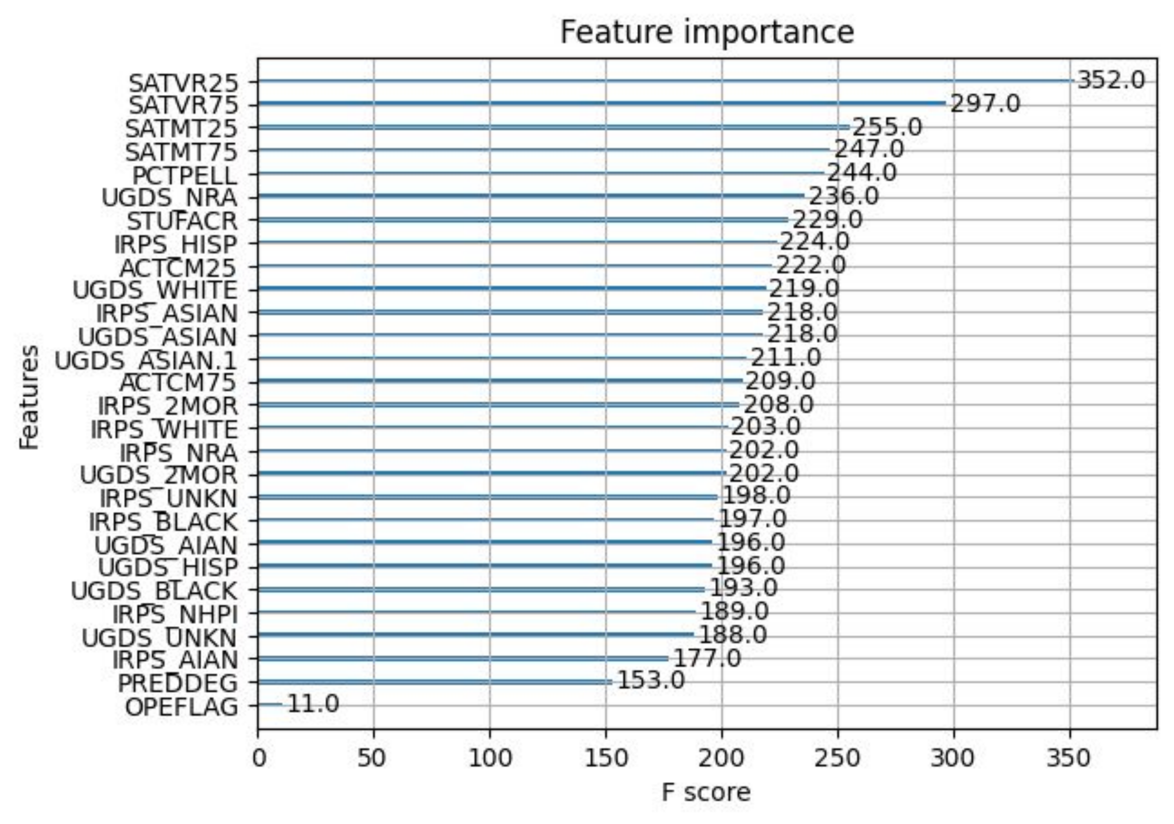
Regression on Iterative Imputation

Model	RMSE	R2
XGBoost	15808.79	0.29
MLP	17404.64	-23.36
Feature Token Transformer	16550.83	-0.01

Manual Cleaning V.S. Imputation

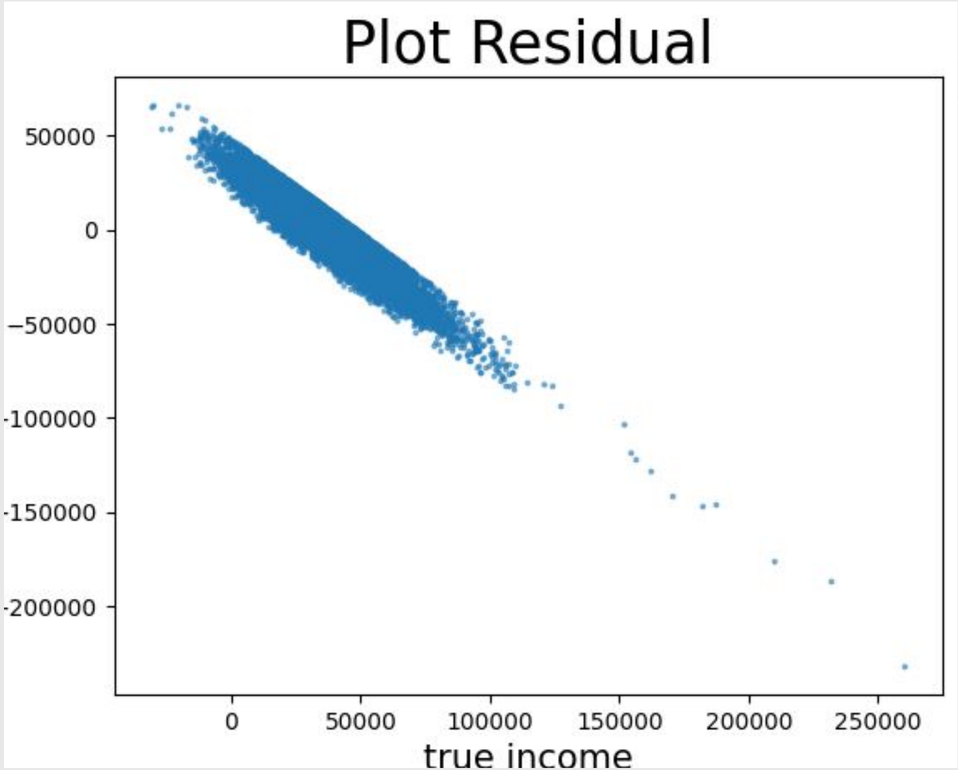
Data	RMSE	R2
Manual	7275.65 (FT Transformer)	0.2 (XGBoost)
Simple	13380.84(MLP)	0.3 (XGBoost)
KNN	13649.53 (XGBoost)	0.55 (XGBoost)
Iterative	15808.79 (XGBoost)	0.29 (XGBoost)

XGboost



Feature importance
SATVR25
SATVR75
SATMT25
SATMT75
PCTPELL
UGDS_NRA
STUFACR
IRPS_HISP
ACTCM25
UGDS_WHITE
IRPS_ASIAN
UGDS_ASIAN
UGDS_ASIAN.1
ACTCM75
IRPS_2MOR
IRPS_WHITE
IRPS_NRA
UGDS_2MOR
IRPS_UNKN
IRPS_BLACK
UGDS_AIAN
UGDS_HISP
UGDS_BLACK
IRPS_NHPI
UGDS_UNKN
IRPS_AIAN
PREDDEG
OPEFLAG
352.0
297.0
255.0
247.0
244.0
236.0
229.0
224.0
222.0
219.0
218.0
218.0
211.0
209.0
208.0
203.0
202.0
202.0
198.0
197.0
196.0
196.0
193.0
189.0
188.0
177.0
153.0
11.0
Features
0
50
100
150
200
250
300
350
F score

FT Transformer



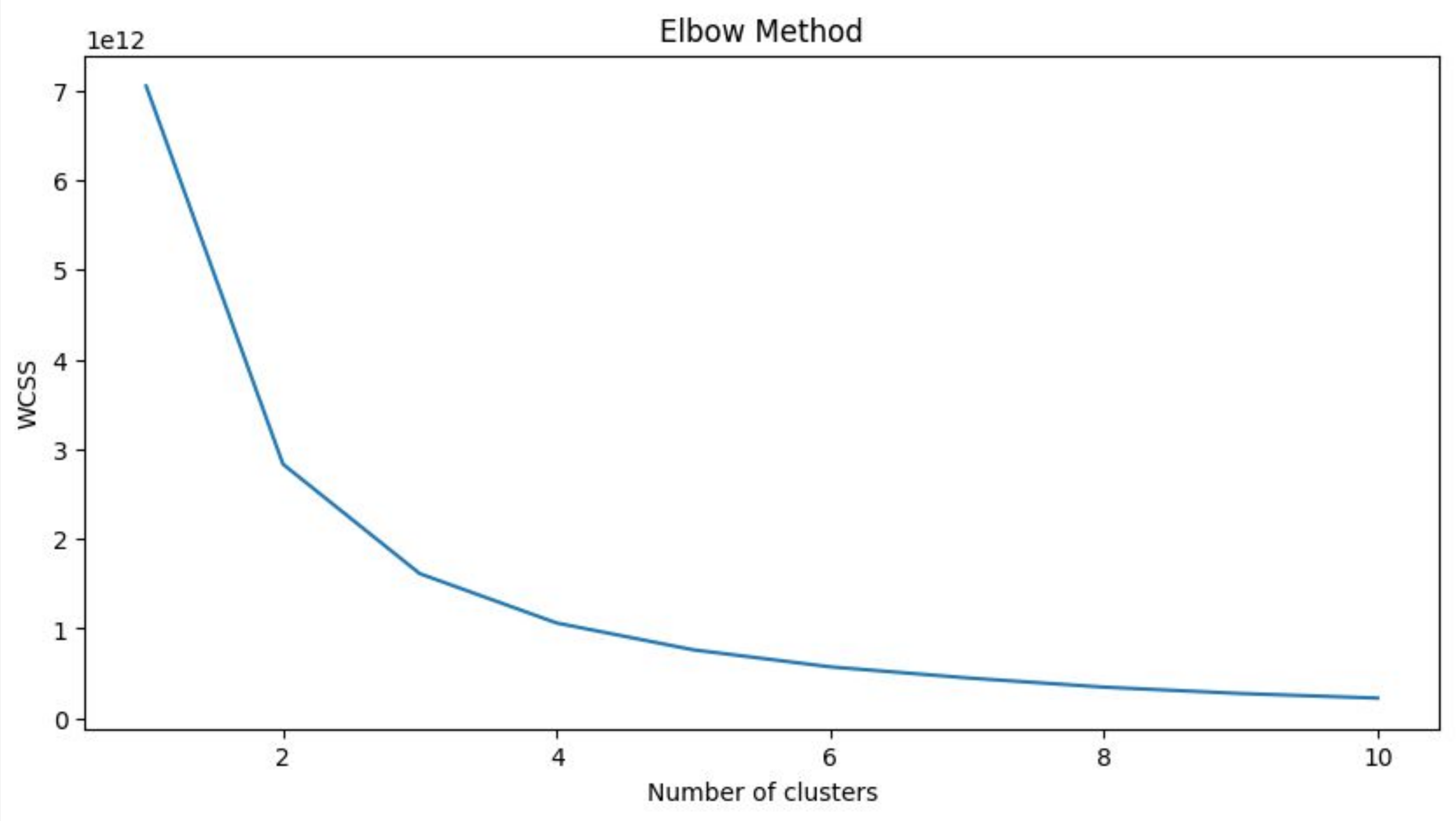
Cluster and Classify

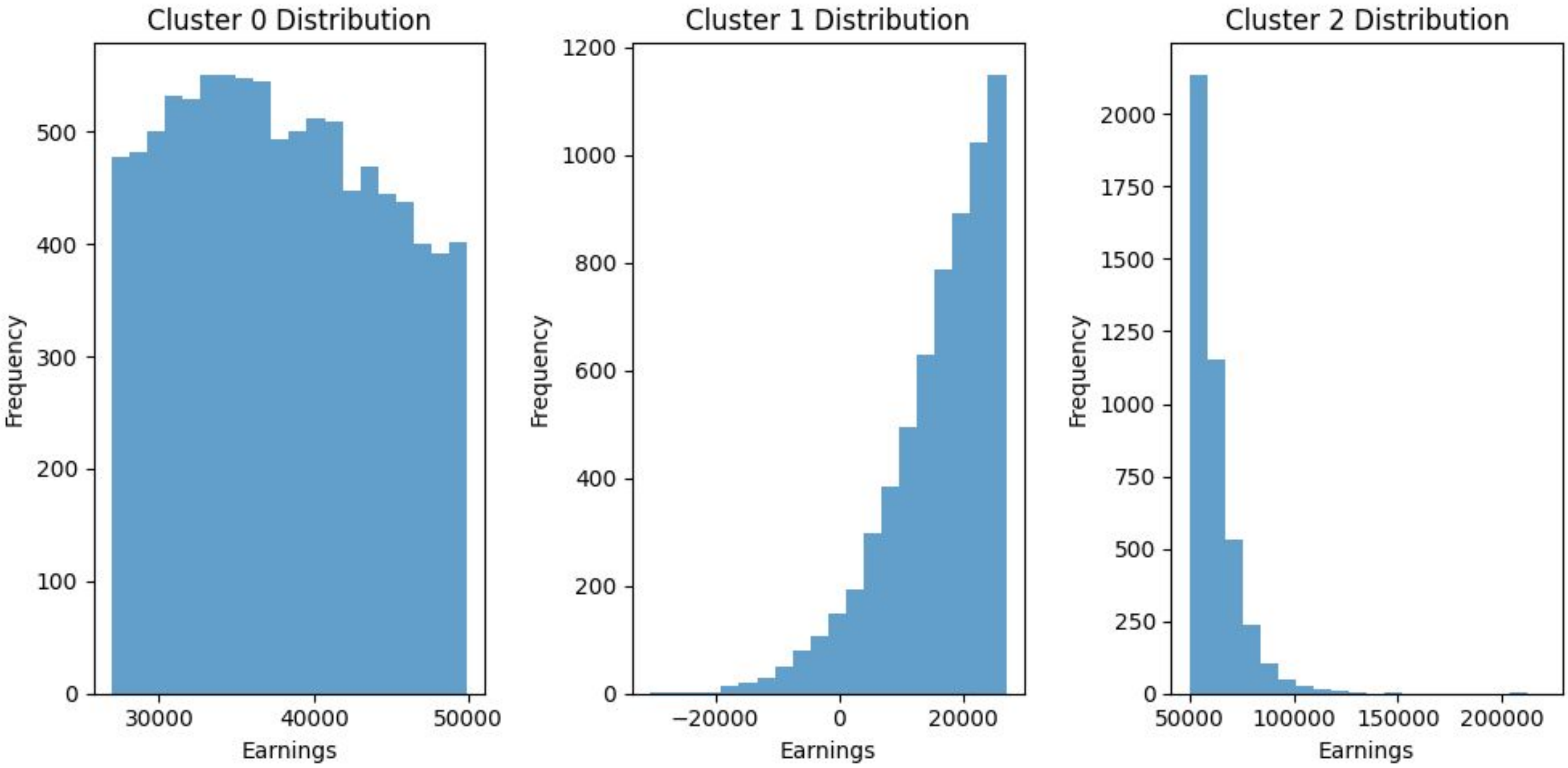
MODEL:

- Random Forest
- XGBoost
- MLP

Data:

- Use KNN to cluster the universities by the income
- Find the number of clusters to cluster
- Relabel the instances with cluster index and classify


Elbow Method
1e12
7
6
5
4
3
2
1
0
WCSS
2
4
6
8
10
Number of clusters


Cluster 0 Distribution
Cluster 1 Distribution
Cluster 2 Distribution
Frequency
Earnings
0
100
200
300
400
500
30000
40000
50000
0
200
400
600
800
1000
1200
−20000
0
20000
0
250
500
750
1000
1250
1500
1750
2000
50000
100000
150000
200000

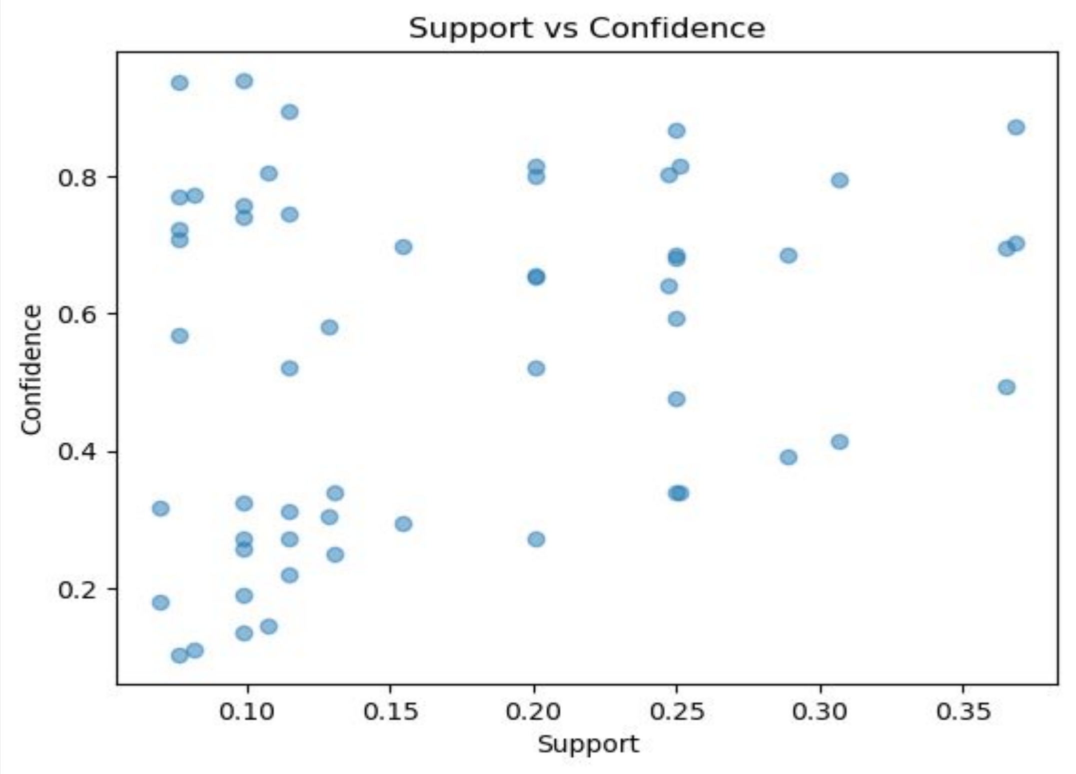
Cluster: Manual Data

Model	Accuracy	Precision	Recall
XGBoost	0.99	1	1
MLP	0.98	0.98	0.98
Random Forest	0.99	1	1

Cluster: Imputed Data

Model	Accuracy	Precision	Recall
XGBoost	0.99	1	1
MLP	0.98	0.98	0.98
Random Forest	0.99	1	1

Recommender System 06



Discussion and Future Work 07

Discussion:

Current Dataset is hard for training:

- Fairly large loss on Income Prediction
- Almost accurate prediction after clustering
- Either the task is too difficult or too easy

By Model:

SAT is one of the main factor:

- If SAT is larger, then higher income in the future

Future Work:

- More robust dataset for this task
- Drop features to train on models
- Add more clusters to get more explainable information
- Investigate on table understanding tasks

References 08

References:
【1】Frontiers in Education. "Developing and Evaluating a University Recommender System." Accessed [04/7/2024]. https://www.frontiersin.org/articles/10.3389/feduc.2020.00135/full.
【2】MDPI. "A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data." Accessed [04/07/2024]. https://www.mdpi.com/2076-3417/11/4/1445.
【3】Sharma, C. "University Recommender." GitHub repository. Accessed [04/07/2024]. https://github.com/chinmaysharmacs10/University_Recommender/tree/master.
【4】Shrooq Algarni, Frederick Sheldon, "Systematic Review of Recommendation Systems for Course Selection", Mach. Learn. Knowl. Extr., 2023, vol. 5, no. 2, pp. 560-596. Available online: https://doi.org/10.3390/make5020033
【5】Hongde Zhou, Fei Xiong, Hongshu Chen, "A Comprehensive Survey of Recommender Systems Based on Deep Learning", Applied Sciences, 2023, vol. 13, no. 20, 11378. Available online: https://doi.org/10.3390/app132011378