Income Prediction for Students

BY: AUDREY EMERIBE

INTRODUCTION

Institutions of higher education usually aim to boost their alumni outcomes after graduation by collecting data from their alumni and identifying patterns seen from those who achieved higher outcomes and those who did not. The results can help guide stakeholders to support future alumni as well as lend guidance and direction in the search and decisions of prospective students.

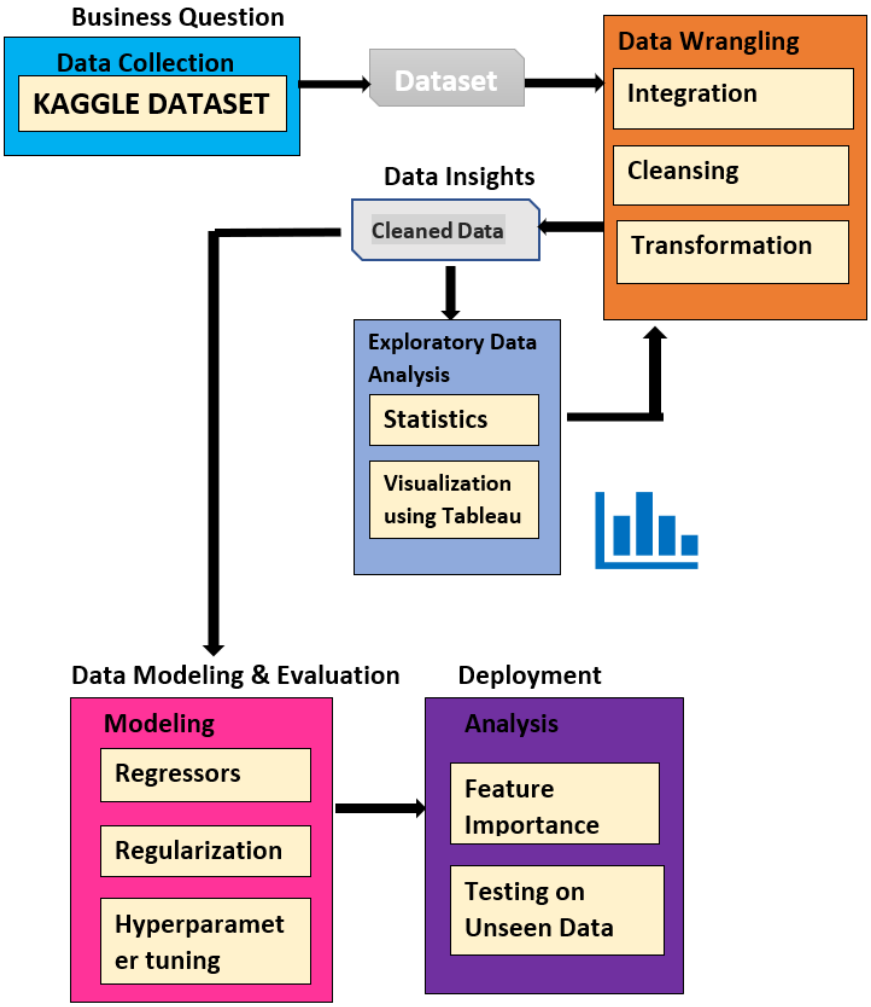
This report shows the study and measurement of the income of alumni students following their degree enrollment, graduation and job market value as a product of their educational background. These studies are beneficial to the evaluation of a higher institution's efficiency and also the support of institutional planning and future student's achievement.

METHODOLOGY

DATA COLLECTION:

This study analyses data collected from the Kaggle called the Microsoft Professional Capstone Dataset made available by Harsh Sharma. The dataset was downloaded as a folder container 3 CSV files (test_values.csv, train_values.csv and train_labels.csv). The data points represented a United States institution of higher education in a specific year while the columns represented various academic program of study, different admission scores (such

as ACT and SAT scores), completion rate for different degrees, educational cost (tuition and fees), different degrees awarded by institutions, different institutional characteristics, ownership and location, family background of students and enrollment rate for institutions. There were 297 columns in the dataset and 26, 299 rows. The target variable being measured is **income**.



FLOWCHART OF THE METHODOLOGY

DATA WRANGLING

The dataset was read into the notebook and analyzed for duplicated rows and columns. The duplicates were dropped. Also, columns with over 70% of missing values were removed from the dataset and the columns with lesser missing values were filled with the median value for each column. Columns that had over 70% of zeros were also dropped as they could influence the model performance during model fitting. Categorical columns were also converted to numerical columns and at the end of this process, we were left wit 73 columns and 26,186 rows. The clean dataset was saved as a csv for data visualization in tableau as well as split for model fitting. This is a regression problem and as such the models to be fitted will all be regression models.

DATA INSIGHTS

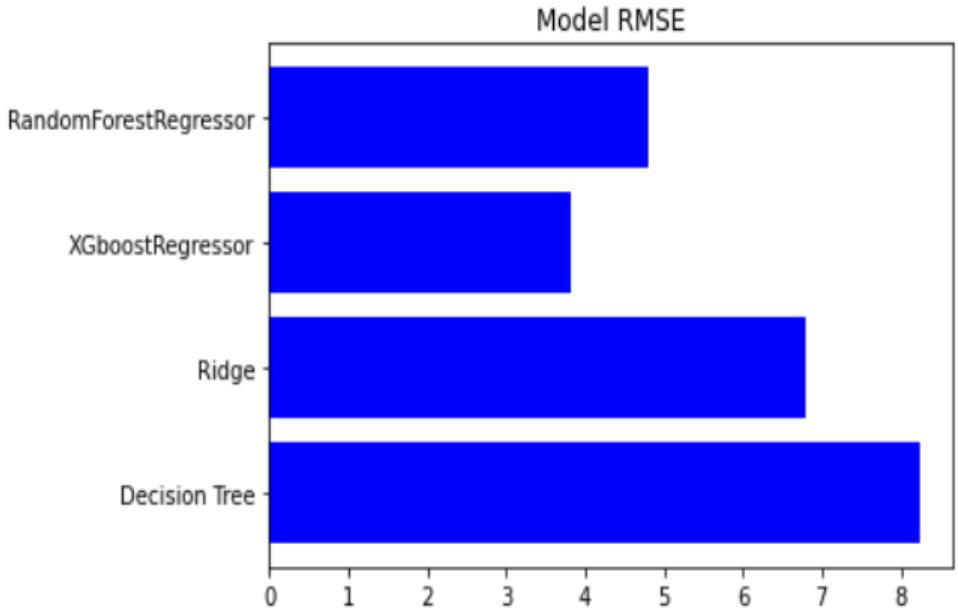
From the visualizations generated, we see patterns where income is influenced by the school degree awarded, location or region of the school, school ownership (whether private for profit, non profit or public institution) and instructional expenditure for each school.

DATA MODELING

To evaluate the dataset, different models were fitted to the train set of the data. Because there was no target variable for the test set, the train set was split using the Scikit Sklearn train-test split

with a test of 0.2. Using Ridge regression, decision tree regression, extreme boosting regressor model and Random forest regressor model, we fitted the train set and predicted both the train set and test set, evaluating the Root Mean Squared Error (RMSE). The RMSE is used to judge the performance of a regression model. It is the standard deviation of the residuals (prediction errors) which shows how concentrated the data is around the best fit. The lower the RMSE, the better the prediction of the model.

After hyperparameter tuning and grid search, we settled on the model (XGBoost Regressor model) with the best parameters which gave the lowest RMSE of 3.81.



MODEL PERFORMANCE AFTER HYPERPARAMETER TUNING

DEPLOYMENT

Settling on the model which gave the lowest loss function, we used the parameters to fit and predict the train set. Using Shapley values, we also got better insights into the major features of the dataset that had an impact on the prediction of the target variable. Shapley value is the average marginal contribution of a feature value across all possible coalition. The Shapley value is the feature contribution to the prediction. The top 5 features that influence the income predictions are: school_degrees_awarded_predominant recorded, school_faculty_salary, academics_program_percentage health, school_degrees_awarded_highest_Graduate_degree and school_ownership_Private for profit.

CONCLUSION AND NEXT STEPS

With so much advancement in analytics, there have been improvements in educational learning process and evaluation of higher institutional efficiency. With this study, we were able to predict income of students with an accuracy of 88% and also identify features that had a strong relationship with student's income.

For future work, it will be a great endeavor to search for better data with lesser missing values and also better features to aid the prediction of income. Also, the year which wasn't given precisely would have been a good factor to evaluate the changing income range and prediction.