Model to estimate the earnings potential of students

Capstone Project – 1

By

Harsh Singh

Springboard: Data Science Career Track

Background

- Factors affecting an individual's decision to join a college:
  - cost of education
  - programs offered
  - estimated student debt incurred
  - earning potential after the completion of degree
  - Others
- Based on the personal preferences and financial condition, future students can have several colleges to choose from

Objectives

- Provide an exploratory data analysis on different variables and their effect on the earnings of students graduating from various colleges. Also, present some Inferential statistics and initial recommendations.

- To predict the earnings of a student graduating from a different colleges using college score data provided by Federal government using different models and compare their performances.

Potential beneficiaries/ Clients:

- Students: to make better decisions
- Colleges: to improve on their spending in the right direction
- Policymakers: policymakers to get better information to provide colleges with more support

Methodology:

- Data Wrangling
- Exploratory Data Analysis
- Inferential Statistics
- Model development

Data Wrangling

- Ingesting the data: Using urls
  - Most-Recent-Cohorts-All-Data-Elements.csv - This file contains all the data needed for developing the project.
  - CollegeScorecardDataDictionary.xlsx – This file consists of details about the variables in the above csv file.
- Generating a dataframe:
  - converted into a pandas dataframe (pd.read_csv)
  - Replace null values represented as ‘PrivacySupressed’ into ‘na’
- Checking Dataframe
  - Using dataframe.head(), dataframe.info(), and dataframe.describe()
  - dataset consists of 7593 rows and 1805 columns

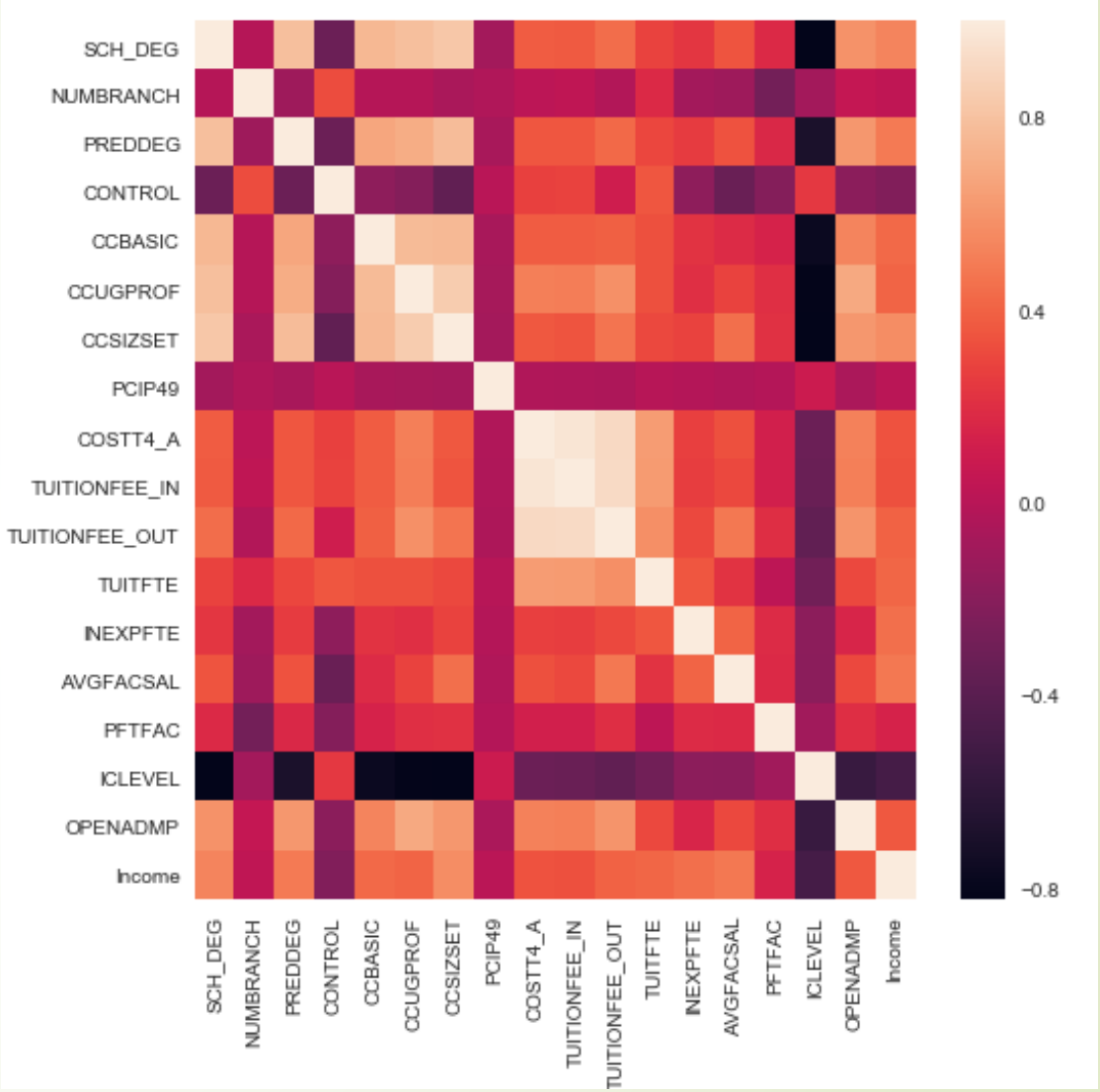
Data Wrangling

- Selecting relevant categories:
  - Out of 10 main categories 5 most important were retained
  - After dropping 5 categories: 159 features left
- Feature Engineering:
  - new column showing the region of the college using the first digit obtained from the zip code
- Treating NULL values:
  - removed all the columns that had more than 40% of NULL values
  - rest of columns we used their medians
- RandomForestRegressor:
  - After fitting the model we had 18 features consisting of continuous and categorical variables

Check collinearity between independent variables

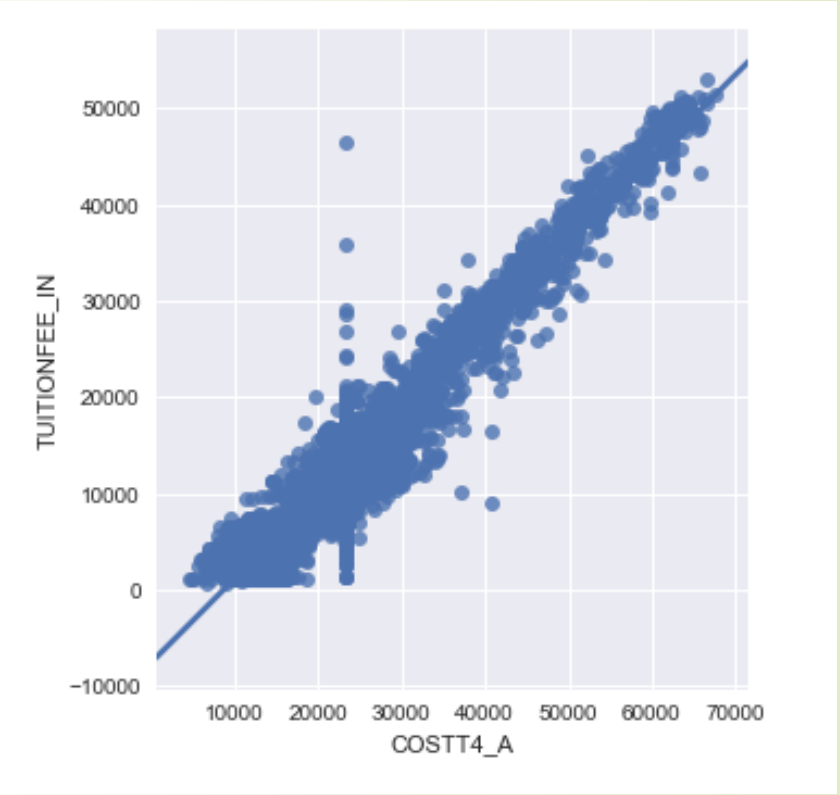
- Heatmap was plotted to visualize the correlation between all the variables.

- **Conclusion:** The plot showed strong relationship between: Tuition fee and cost of attendance



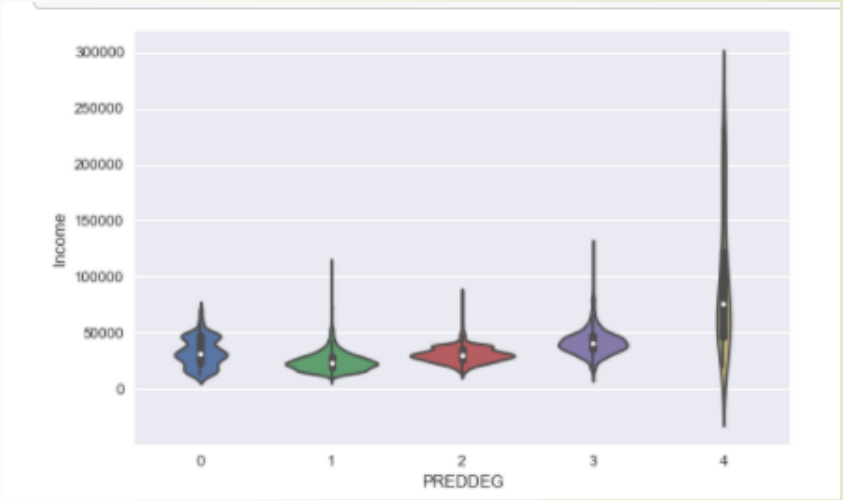
Collinearity between Tuition fee and cost of attendance

- Bootstrap method was used to test the significance of collinearity
- The p-value for the test was 0.0, this confirmed that the cost of attendance and tuition fee are statistically strongly correlated
- **Conclusion:** since the estimated value of correlation coefficient (0.97) is high we can say that it is also practically significant.



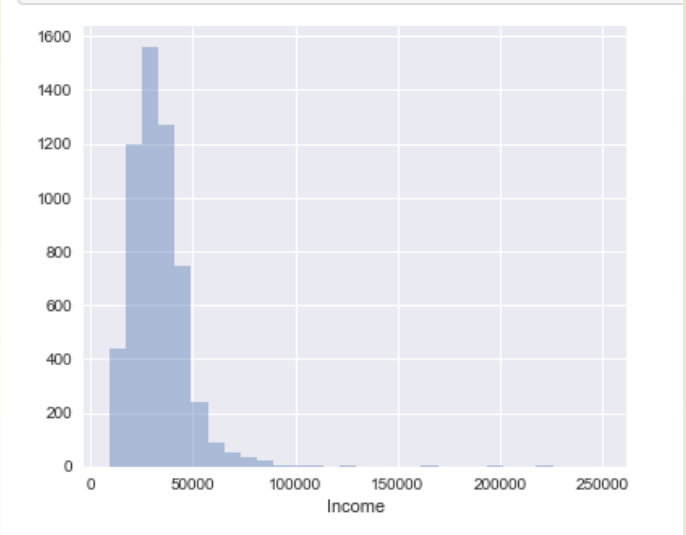
Compare mean income of students graduating from colleges that gives predominantly graduate degrees against other colleges

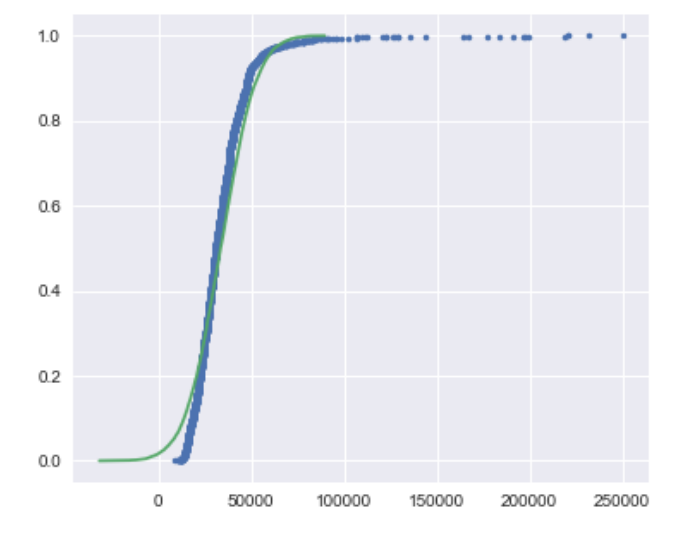
- The empirical difference between the means of the groups was around $ 61,684, with mean income of colleges giving predominantly graduate degrees around $94, 447
- Bootstrap method along with the t-test using stats package were performed to check if the difference was statistically significant.
- **Conclusion:** Both the tests suggested that the salaries of the student from colleges offering graduate degrees were significantly higher.



Check the normality of the Income data

- Histograms and empirical cumulative distribution functions were plotted for the Income data
- The data at lower and higher extremes (could be outliers) deviate the plot from being normally distributed
- Tested the normality of the income data using chi-square test.
- **Conclusion:** Data is not normally distributed
- Take the log of data to make it normal





Model Development

- Random Forest Regressor
  - Accuracy
  - Can easily handle large number of variables
- Elastic Net
  - Overcomes the limitations of both lasso and ridge regression
- Compare the performances of both models during training and testing

Random Forest Regressor

- Split into training (70%) and testing (30%) dataset
- Training: Grid Search for Hyperparameter Tuning
  - Cross-validation: 5 partitions
  - max_depth: list(range(1, 10))
  - min_sample_split: [50, 100, 200]
  - n_estimators: list(range(1, 5))
  - max_features: list(range(2, 18))
- Testing on rest of the 30% data
- Repeat the process using pipeline and scaling the income data

Elastic Net

- Split into training (70%) and testing (30%) dataset
- Training: Grid Search for Hyperparameter Tuning
  - Cross-validation: 5 partitions
  - l1_ratio: linspace(0, 1, 30)
- Testing on rest of the 30% data
- Repeat the process using pipeline and scaling the income data

Results:

	Random Forest Regressor		Elastic Net	
	Income	Log-Income	Income	Log-Income
R-square	0.613	0.647	0.522	0.522
RMSE	10124.948	9679.449	126844080.628	126861256.234

- Random Forest Regressor was able to develop better model with higher R-square and much lower RMSE
- Standardizing Income data by taking log further improve the performance of Random Forest Regressor model.

THANK YOU