**Model to estimate the earnings potential of students**

**Introduction:**

Taking a decision to attend a college is one of the most difficult decisions an individual has to make in their lifetime. Several factors that dictate this decision include cost of education, programs offered, estimated student debt incurred, earning potential after the completion of degree and others. Based on the personal preferences and financial condition, future students can have several colleges to choose from. There are websites that provide information on rankings of the colleges based on these criterions.

In this project, I want to focus on developing a model to estimate the earning potential of college graduates 10 years after completing their degrees based on the college they graduate from. I will also compare the salaries estimated by my model using the college scorecard data against the salaries provided by other available sources (e.g. websites).

Several variables (1805) are available from 1996 to 2016 for institutions of the United States including cost, graduation rate, debt, programs, SAT scores, repayment, earnings after school and others variables (<https://collegescorecard.ed.gov/data/>). A report from federal government compared the effect of these variables on the earnings of the students using aggregate and individual level regression approaches (Executive office of the President of the United States, 2015). In the project they also explored the effect of these variables on the earning potentials of the students.

Overall the objective of this capstone project is to:

1. 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.
2. 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.

**Business problem:** *To predict the earnings of students graduating from different colleges.*

**Dataset availability:**

1. College scorecard dataset in the form of csv files downloaded from: <https://collegescorecard.ed.gov/data/>
2. Earnings of various alumni from different colleges from websites:

e.g. <https://www.payscale.com/>

**Potential beneficiaries/ Clients:**

The model developed during this capstone project can help students to make better decisions while choosing a college that can maximize their earning potential, colleges to improve on their spending in the right direction to help their future students to be more successful, policymakers to get better information to provide colleges with more support.

**Overall Methodology:**

All the codes for this project will be developed using python language in Jupyter notebook. I will start the project with data ingestion from the csv files, cleaning the data using data wrangling techniques. Next step, would be to perform exploratory analysis by analyzing the effect of independent variables on the dependent variable (median earnings of student). Calculate initial statistical inferences from the data samples.

Divide the data into training and testing datasets. Use cross validation, grid search on the training dataset to develop models for estimating the earnings of an individual based on the independent variables. Test the performance of different models in predicting the response variable using the rest of the testing dataset. Score and compare the models based on their performance during training and testing phases.

Compare the results from the best model against other dataset obtained from other websites.

**Data wrangling:**

Two datasets were imported from the website (<https://collegescorecard.ed.gov/data/>) using their urls:

1. Most-Recent-Cohorts-All-Data-Elements.csv - This file contains all the data needed for developing the project.
2. CollegeScorecardDataDictionary.xlsx – This file consists of details about the variables in the above csv file.

The steps taken to clean the data are as follows:

***Generating a dataframe:*** The csv data was downloaded from the url stated above and was converted into a pandas dataframe (pd.read\_csv). It was noticed that null values in several columns were represented as ‘PrivacySuppressed’. These values were converted into “na” values.

***Checking dataframe:*** The dataframe consisting of relevant categories was checked using commands - dataframe.head(), dataframe.info(), and dataframe.describe(). The dataset consists of 7593 rows and 1805 columns.

***Selecting relevant categories***: The data in the main csv file can be divided into 10 subcategories (e.g. school, admissions, academics, student, cost). We used the data dictionary to get the unique categories. Out of these we found 5 categories that were relevant for our analysis. For instance, we discarded the categories that focused on demographics of the student such as ethnicity, gender etc. After dropping these columns, we were left with 159 variables.

***Transforming Zipcode into categorical data:*** In order to divide the colleges into different regions, their zipcodes were used. We generated a new column showing the region of the college using the first digit obtained from the zip code. (df[‘Zip’.map(lambda x: int(str(x)[:1])).

***Treating NULL values:*** Percentage of NULL values in each column were calculated (df.isnull().sum() / df.shape[0] \*100). It was noticed that several columns had more than 40% of NULL values. So, we removed all the columns that had more than 40% of NULL values. This technique reduced the number of columns to 65. Further to treat the NULL values in the rest of columns we used their means (df.fillna (df.mean(), inplace=True)

It was noticed that 38 of the 65 columns represent the percentage of students in a certain program such as agriculture, computer, engineering including others. Some of the programs are not that common, resulting in several zeroes. The columns with more than 90% of zeroes were removed. Finally, we had 28 columns.

***RandomForestRegressor:*** RandomForestRegressor was imported from sklearn.ensemble and SelectFromModel from sklearn.feature\_selection to reduce the number of variables. After fitting the model we had 18 variables consisting of continuous and categorical variables.

***Normalization and Outliers:*** Income was normalized between 0 and 1. The boxplots were plotted for variables to identify outliers.

**EDA and Inferential Statistics:**

1. *Check collinearity between independent variables*

A heatmap was plotted to visualize the correlation between all the variables. The plot showed that there is strong relationship between tuition fee and cost of attendance. In order to further substantiate the relationship, we used scatter plot (sns.plot). Finally, we used bootstrap method to check the statistical significance of the collinearity between tuition fee and the cost of attendance.

* Correlation coefficient between the two variables was estimated using numpy package (np.corrcoef(x,y)). The estimated value was 0.97
* We also tested the significance (for alpha =0.05) of the value using bootstrap method:

1. Null hypothesis: Cost of attendance is completely uncorrelated to tuition fee.
2. Alternate hypothesis: Cost of attendance and tuition fee are correlated.

* The p-value for the test was 0.0, this confirmed that the cost of attendance and tuition fee are statistically strongly correlated. Also, since the estimated value of correlation coefficient is high we can say that it is also practically significant.

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

Violin plot was generated to visualize the spread of income for students from different colleges based on the type of degree predominantly awarded by the college (1. Certificate, 2. Associate, 3. Bachelor, 4. Graduate). The graph showed that mean income was higher for colleges that awarded graduate degrees. We further checked it using statistical analysis

* We separated the colleges in two groups: colleges that predominantly give graduate degrees and colleges that do not predominantly give graduate level degrees.
* 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.

1. Null hypothesis: students from graduate level colleges have similar salaries as the student from other colleges
2. Alternate hypothesis: students from graduate level colleges have salaries significantly higher as compared to the students from other colleges.

* Both the tests suggested that the salaries of the student from colleges offering graduate degrees were significantly higher.

1. *Check the normality of the Income data*

Histograms and empirical cumulative distribution functions were plotted for the Income data. Also, to confirm the normality we performed chi-square test.

* Both the plots suggested that, while the data in the middle looks close to normally distributed, the data at lower and higher extremes (could be outliers) deviate the plot from being normally distributed.
* We also tested the normality of the income data using chi-square test. Finally, we concluded that the data is not normally distributed.
* We also checked the normality of log transformed Income data, and it was observed that the log transformed income data is normally distributed.

**Recommendations for further analysis and building the model:**

* We can eliminate one of the variables, either tuition fee or the cost of attendance.
* We can divide the dataset into two groups based on if they awarded the graduate degrees or not.
* We can use log transformations of the income data for further analysis.
* We should also try centering and scaling all the variables to improve the model performance.

**Model Development**

Random Forest Regressor and Elastic Net models were developed to predict the income of student 10 years after their graduation. Data was split into training (70%) and testing (30%) sets using train\_test\_split function from model\_selection module of Sklearn package. In order to test the effect of scaling the data we used both Income and log of Income as the response variables for developing the model.

Random forest was chosen since it is one of the most accurate algorithms available. It can handle large number of input variables and it also return the importance of the variables. Elastic Net was used since it solves the limitation of both ridge and lasso regression.

**Random Forest Regressor (using Income raw income data)**

***Training:*** Hyperparameters (max\_depth (from 1 to 10 using list); min\_sample\_split (50, 100, 200); n\_estimators (1 to 5 using list); max\_features (from 2 to 18 using list)) were tuned using Grid Search with cross validation using 5 partitions. The model was tuned using the training data, result of the hyperparameter tuning was:

Tuned Random Forest Regressor Parameters: {'max\_depth': 9, 'max\_features': 10, 'min\_samples\_split': 50, 'n\_estimators': 4}

Best Score is: 0.623650844066469

***Testing:*** Model was tested on rest of the 30% data. Model predicted the data with good accuracy (R-square: 0.61 and RMSE: 10124.94860160248

**Elastic Net (using Income raw income data)**

***Training:*** Elastic net model was used to predict the income using predictor variables. Since it uses L1 and L2 regularizers. L1\_ratio was tuned on a linspace. The result of Grid Search suggested that best model was obtained using L1 ratio as 1 or in other words using only L1 penalty.

***Testing:*** Model was tested on the test data and the model accuracy was R-square: 0.52 and RMSE: 126844080.62861225.

The result suggested that Elastic Net model did not perform as well as compared to the Random Forest Regressor Model.

**Random Forest Regressor (using log Income income data)**

***Training:*** Model was scaled and fit using pipeline (steps = [('scaler', StandardScaler()), ('regr\_sc' , RandomForestRegressor())). Same hyperparameters were tuned for the model using Grid Search method.

Results of the model accuracy were:

The results of the performances of both Random Forest Regressor and Elastic Net models in predicting income of the students for both raw and log income data are presented in the table. The results suggest that both in terms of R-square and RMSE, Random Forest Regressor model was performing better.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 |

The comparison between Random Forest Regressor and Elastic Net model shows that Random Forest Regressor was able to develop better model with higher R-square and much lower RMSE as compared to Elastic Net model. Standardizing Income data by taking log further improve the performance of Random Forest Regressor model. The biggest disadvantage of random forest could be that it can overfit with noisy data but in our case we can see that the random forest did a good job in predict test data. So, we should be able to generalize it.

Certain biases include, that less data available for schools with fewer number of students. Some institutions report their data at institution level while others at campus levels.

**References:**

Executive office of the President of the United States. (2015). Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education. Retrieved from https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf