IT5006

Fundamentals of Data Analytics

Project Report: US College Scorecard

Group Number: 17

Group Members:

Prayogi Tio [A0255946N]

Yap Wei Xuan [A0183226H]

Ong Jia Sheng [A0250967U]

Lwee Yong Xin, Michael [A0255993L]

Date: 14/04/2023

Introduction

Education is linked to success, as seen in studies where the median salary of university graduates in Singapore is 62% higher than those with only diplomas or A-levels. This raises the question: what factors contribute to the varying wages of university graduates?

In the USA, parental education and income are strong predictors of a child's educational attainment, likely due to the expensive college education system and legacy admissions in elite institutes. This report aims to explore the differentiating factors that contribute to the success of some university graduates, such as student and college backgrounds. Data from the US College Scorecard dataset, which is a publicly available dataset published by the US Department of Education, will be used to examine the above mentioned factors.

Data Description

Sparse data is a challenge in this study, requiring us to choose features with less missing data to avoid false conclusions. We reject anything with over ~35% missing data, unless the feature has a strong correlation. We also consider features based on domain knowledge, despite weak correlation with our success variable.

Due to the large amount of features present in the data, there is a further consolidation in our feature selection by collapsing similar features (e.g. DEBT_ALL_PP_ANY_MEAN, DEBT_ALL_PP_ANY_MDN, which is the median and mean of Parent PLUS loan debt), and choosing the one which has the least missing data.

Lastly, we faced the problem of choosing our success variable and the features. For our success variable, we chose MD_EARN_WNE_P10, median earnings of students working and not enrolled 10 years after entry, which has the least amount of missing data. We assessed the correlation of each feature against our success variable, and would select the feature, condition on it fulfilling our previous requirements.

Data Cleaning and Preparation

We initially attempted to stitch together the data for the whole period which was available to perform our analysis. However, this was unfortunately thwarted due to hardware limitations, and instead we chose filtered down to data from 2013 to 2015 and 2018 to 2019 for our project.

The periods were chosen due to the completeness of the feature data, and success variable relative to the other periods, as well as still being able to run properly on our hardware. Our count of data points is 17609, and the data was split into a train-test split of 80/20. Our initial list consists of 28 features (1 categorical + 27 numerical) in total.

We constructed our data pipeline out of 3 main transformers, the One Hot Encoder to transform categorical features into binary vectors of zeros and ones, Simple Imputer to replace the missing data in any column with the mean value of the column, and Standard Scaler to standardize the data.

Data Exploration and Visualization

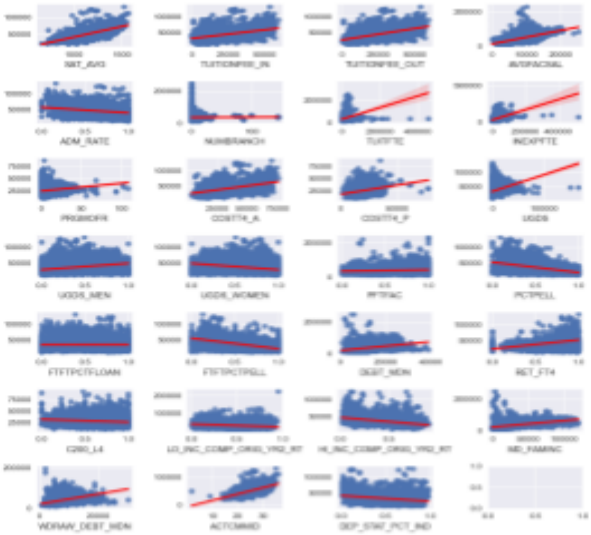

Fig.1 Trend Plots

We computed summary statistics for each feature and plotted a regression line for each feature against the median salary.

After analyzing the regression plots, we have identified several features that show minimal correlation with our target variable. Hence, we have decided to drop these features after building our initial linear regression model.

We then proceeded further to analyze the data using increasingly complex models.

Please refer to the jupyter notebook *"Group17 - Salary Estimator.ipynb"* for the statistics and correlations (*section "Recipe of the regression", step number 2)*.

Data Analysis

In our project, we have utilized multiple models in a stepwise manner to explore various techniques for analyzing the data.

Model 1. Linear Regression Models

We initially fitted all 28 features into a linear regression model and analyzed the P-values, residual plot, and Q-Q plot. We then attempted to remove heteroscedasticity by fitting the log of the target variable into a linear regression. We then dropped moderate to highly collinear features in order to improve the model. We also dropped features with a P-value above 0.05 as there is insufficient evidence to reject the null hypothesis. With little success on linear regression, we moved on to polynomial regression instead. This successfully increased the R-squared by approximately 7%, and also improved mean squared error and root mean squared error.

Model 2. Ensemble Learning Models

As our regression models were incompatible with the data, we explored ensemble learning methods such as Random Forest Regressor and XGBoost. These methods were suitable for

our data, which exhibited general non-linearity and required classification of features by predictive strength. The two methods were chosen to corroborate and double confirm findings.

Model 3. Neural Network Model

We also tried a neural network model with the Adam optimization algorithm, minimizing mean squared loss by selecting coefficients for features. This approach worked well with non-linear data and could optimize the results from the previous model.

Result Discussion and Conclusion

Model 1. Linear Regression Models

Our best regression model is multiple polynomial model (n=2) based on the results below, based on the R-squared value, mean squared error, and root mean squared error.

Model	R-Squared	Mean Square Error	Root Mean Square Error	F-Statistic	P(F-Statistic)
Linear	0.58	94374899.58	9714.67	560.0	0.00
Linear (Log Target Variable)	0.53	104561156.24	10225.51	864.2	0.00
Linear (Dropped several features)	0.55	100239703.03	10011.98	853.3	0.00
Polynomial (n=2)	0.65	77882945.54	8825.13	154.8	0.00
Polynomial (n=3)	0.58	94067277.97	9698.83	49.89	0.00

We observed from our results that the residual plot is funnel-shaped, which suggests there is heteroscedasticity, and our data might not follow a linear model.

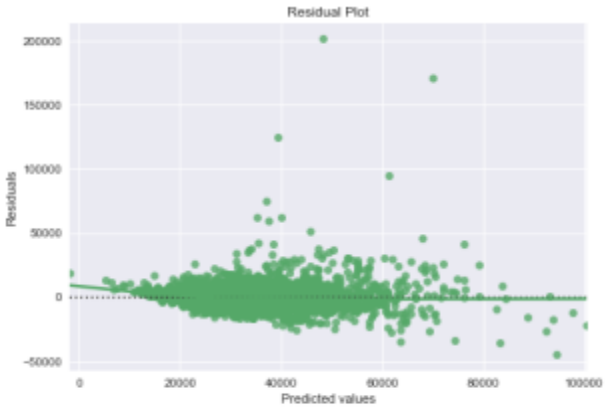

Fig.2 Residual Plot for linear regression

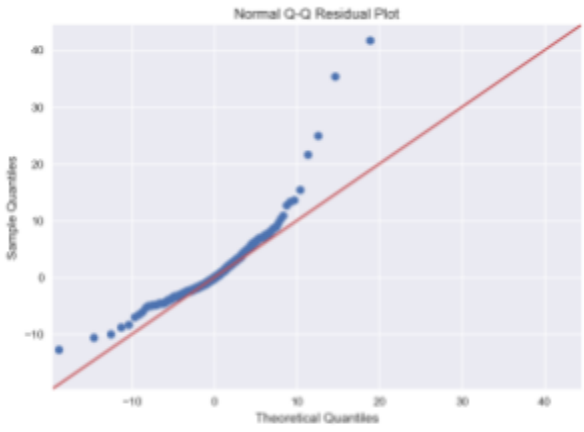

Fig.3 Q-Q Plot

We also performed a QQ-Plot to check whether the residual fits into gaussian assumptions.

We saw that our data did not fit a gaussian model, so we tried fitting polynomial features into a regression, but with limited success. The R-squared, MSE, and RMSE values did improve, but the results were still not satisfactory.

Linear Regression models were unable to provide a good estimate for our data, possibly due to varying underlying distributions for each feature and non-linear data. To confirm this, we also used ensemble methods and neural networks.

Model 2. Ensemble Learning Models

Both ensemble learning methods have performed much better than our regression model.

Model	R-Squared	Mean Square Error	Root Mean Square Error
Random Forest	0.84	34862157.23	5904.42
XGBoost	0.86	30884170.62	5557.35

Ensemble learning methods confirm that our data is non-linear, with XGBoost performing better than Random Forest in all aspects.

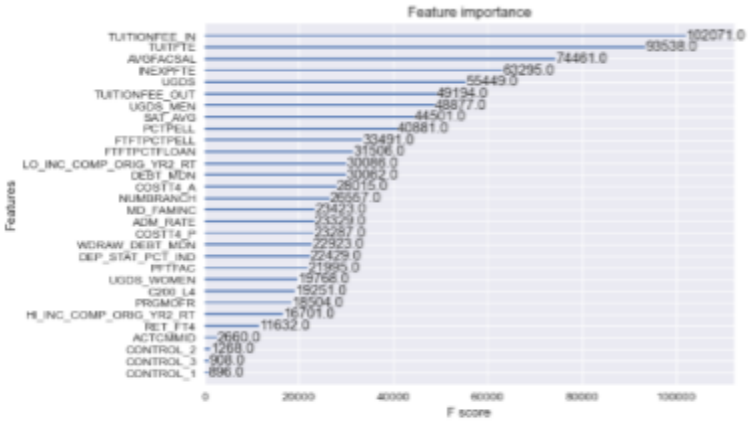

Fig.4: XGBoost Regressor Feature Importance Graph

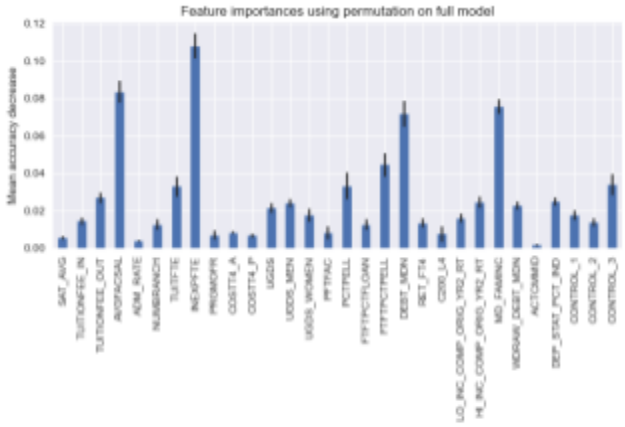

Fig.5: Random Forest Regressor Feature Importance Graph

Key factors such as tuition fees and income, faculty salary, higher expenditures per student and median family income are the top features that explain our success variable. However, the size of the undergraduate cohort also plays an important role, which is surprising.

The regression graph shows an upward trend for the size of the undergraduate cohort against median income. However, this trendline is negatively influenced by selective institutes such as Harvard, Princeton, or MIT with low enrollment sizes and high median income. This suggests a possible synergistic effect, where larger cohort size results in more competition leading to better student quality or a larger alumni network helping students secure better jobs.

Model 3. Neural Network

The neural network was performed initially with the aim of producing a more optimum result as compared to what we have conducted in the ensemble learning methods. However, our R-squared score of 0.80 pales against the 0.86, whereas the mean squared error and root squared error is higher than that of the ensemble learning methods.

Model	R-Squared	Mean Square Error	Root Mean Square Error
Neural Network	0.80	44579200.00	6676.77

Neural network model performed worse than ensemble learning methods, possibly due to lower sample size as neural networks require large data, whereas ensemble learning methods require less.

Conclusion and Limitations

After comparing three approaches, we learn several key facts:

1. Linear regression models may not capture complex relationships between variables, while more advanced models such as ensemble learning are better suited for this task. Linear regression assumes a uniform relationship between all variables, while other models can better handle more intricate relationships.
2. Complex models are not always better, as demonstrated by the comparison between our neural network and ensemble learning models. Other factors, such as data availability, may make one model a better choice than the other.
3. Regression models and neural networks are more prone to overfitting, resulting in less accurate predictions.
4. Mean imputation can result in significant loss in predictive performance for data with missing values. This is evident in the low feature importance of SAT average, despite its high correlation, due to a large amount of missing data for this feature.

We acknowledge that our models might have the following shortcomings or limitations:

1. The accuracy of the imputation method may affect the results, highlighting a potential impact on the outcomes.
2. The absence of domain expertise may limit our ability to use appropriate techniques for data cleaning and preprocessing.
3. The model construction has been limited by memory errors caused by hardware limitations, resulting in a reduced amount of data being processed.
4. Lack of knowledge in more advanced techniques which we can utilize.

Kaggle Competition **(Best Submission: IT5006_Grp17_Kaggle_Best_Model_XGBoost_Regressor)**

Data Preprocessing

The training data set had outliers that were identified by falling outside of the upper limit (75th Quartile + InterQuartileRange) or lower limit (25th Quartile - InterQuartileRange) ranges, and were capped at their respective upper or lower limits. Additionally, feature engineering was used to combine highly correlated features in the original dataset. Finally, the StandardScaler class was used to normalize both the training and test data sets.

Modeling Approach and Selection of Model

XGBRegressor (**0.88**) outperforms RandomForest (**0.87**) and Neural Network (**0.84**) in prediction quality, and is used to evaluate the impact of Outlier Capping and Feature Engineering.

Models were optimized using GridSearchCV class from SkLearn library with at least 2 runs, narrowing parameter ranges for each subsequent run. Results are presented in a table below.

Model	XGBRegressor (With Outlier Capping)	XGBRegressor (With Feature Engineering)	XGBRegressor (With Outlier Capping + Feature Engineering)	XGBRegressor (Without Outlier Capping / Feature Engineering)
Parameters	**n_estimators**=400 **subsample** = 0.65 **max_depth** = 25 **learning_rate** = 0.06, **colsample_bytree** = 0.9	**n_estimators**=500 **subsample** = 0.6, **max_depth** = 25 **learning_rate** = 0.05 **colsample_bytree** = 0.75	**n_estimators**=400 **subsample** = 0.6, **max_depth** = 23, **learning_rate** = 0.06, **colsample_bytree** = 0.8	**n_estimators**=500 **subsample** = 0.5 **max_depth** = 18 **learning_rate** = 0.05 **colsample_bytree** = 1
Kaggle Score	**0.87679**	**0.87587**	**0.87180**	**0.88243**

Evaluation

XGBRegressor performs better due to its Gradient Boosting iterative method, while Neural Network requires a large data size and regularization can reduce predictive quality. Outlier Capping and Feature Engineering led to slightly lower scores.

Outlier capping and feature engineering resulted in slightly lower scores, possibly due to the small size of our dataset (~2364 data points), where outliers could be statistically significant and the merging of features might increase bias. Hence, XGBRegressor without these techniques performed the best.

Additional NoteBook References: IT5006_Grp17_Kaggle_RandomForestRegressor, IT5006_Grp17_Kaggle_NeuralNetwork, IT5006_Grp17_Kaggle_Capping_Outliers_XGBoost_Regressor, IT5006_Grp17_Kaggle_Feature_Engineering_XGBoost_Regressor, IT5006_Grp17_Kaggle_Feature_Engineering_and_Capping_Outliers_XGBoost_Regressor,

Reference List

- *Data Home: College Scorecard* (2022) *Data Home | College Scorecard*. U.S. Department of Education. Available at: https://collegescorecard.ed.gov/data/ (Accessed: April 14, 2023).
- Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
- Sachdev, H.S. (2020) *Choosing number of hidden layers and number of hidden neurons in neural networks*, *LinkedIn*. Harpreet Singh Sachdev. Available at: https://www.linkedin.com/pulse/choosing-number-hidden-layers-neurons-neural-networks-sachdev/ (Accessed: April 14, 2023).
- Brownlee, J. (2023) *Pytorch tutorial: How to develop Deep Learning models with python*, *MachineLearningMastery.com*. Jason Brownlee. Available at: https://machinelearningmastery.com/pytorch-tutorial-develop-deep-learning-models/ (Accessed: April 14, 2023).
- Ellis, C. (2022) *When to use random forests*, *Crunching the Data*. Christina Ellis. Available at: https://crunchingthedata.com/when-to-use-random-forests/ (Accessed: April 14, 2023).
- Brownlee, J. (2021) *Gentle introduction to the adam optimization algorithm for deep learning*, *MachineLearningMastery.com*. Jason Brownlee. Available at: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/ (Accessed: April 14, 2023).
- Kingma, D.P. and Ba, J. (2017) *Adam: A method for stochastic optimization*, *arXiv.org*. Diederik P. Kingma, Jimmy Ba. Available at: https://arxiv.org/abs/1412.6980 (Accessed: April 14, 2023).
- Cue (2022) *University Grads' median pay is $4.2K, double the $2K of those with ITE, secondary education: Study*, *The Straits Times*. Theresa Tan. Available at: https://www.straitstimes.com/singapore/community/university-grads-median-pay-is-42k-double-the-2k-of-those-with-ite-secondary-education-study (Accessed: April 14, 2023).
- Autor, D. H. “Skills, Education, and the Rise of Earnings Inequality Among the ‘Other 99 Percent.’” Science 344, no. 6186 (May 22, 2014): 843–851.
- Warren, K. (2018) *Here's what college costs in 28 countries around the world*, *Insider*. Insider. Available at: https://www.insider.com/cost-of-college-countries-around-the-world-2018-6#united-states-8202-28 (Accessed: April 14, 2023).
- Saul, S. (2022) *Elite Colleges' quiet fight to favor alumni children*, *The New York Times*. The New York Times. Available at: https://www.nytimes.com/2022/07/13/us/legacy-admissions-colleges-universities.html (Accessed: April 14, 2023).
- *Captum · Model Interpretability for Pytorch* (no date) *Captum*. Facebook Open Source. Available at: https://captum.ai/ (Accessed: April 14, 2023).