**PREDICTING UNIVERSITY GRADUATION RATES: A COMPARITIVE ANALYSIS OF MACHINE LEARNING REGRESSION MODELS**

**THRUSHWANTH KAKUTURU**

**KO67574**

**GRADUATE STUDENT**

**MASTER OF PROFESSIONAL STUDIES IN DATA SCIENCE**

**UNIVERSITY OF MARYLAND BALTIMORE COUNTY**

**INDEPENDENT STUDY**

**MASOUD SOROUSH**

**PROBLEM STATEMENT:**

The landscape of higher education is under scrutiny for institutional effectiveness and student success, with graduation rates serving as pivotal benchmarks. This study, part of the Master of Professional Studies in Data Science program at the University of Maryland Baltimore County, aims to construct machine learning and deep learning models predicting graduation rates at four-year institutions. Focusing on completion rates within specified timeframes, the research employs the College Scorecard database, offering a comprehensive view of institutional data. The primary objectives are twofold: enhance predictive capabilities for educational outcomes and provide stakeholders with a toolset for informed decision-making. By deploying regression models, the study seeks to discern patterns within the data, appraise predictive prowess, and offer insights into the most efficacious methodologies for predicting graduation rates. This research not only contributes to academic understanding but also empowers stakeholders, including policymakers and educational institutions, with actionable insights to improve institutional effectiveness and student success.

**OBJECTIVES:**

* Enhance predictive capabilities for graduation rates at four-year institutions, focusing on "C100\_4," "C150\_4," and "C200\_4" variables that represent completion within 100%, 150%, and 200% of normal time, using machine learning models.
* Identify the most prominent features that influence graduation rates.
* Assess and compare regression models to identify the most effective methodologies for predicting graduation rates, guiding future analyses in higher education.
* Empower policymakers and institutions with actionable insights for informed decision-making to improve student success.

**DATA SOURCES:**

* College Scorecard Website: https://collegescorecard.ed.gov/
* College Scorecard Data Web Page: https://collegescorecard.ed.gov/data
* Documentation: <https://collegescorecard.ed.gov/assets/InstitutionDataDocumentation.pdf>
* Dataset: <https://ed-public-download.app.cloud.gov/downloads/CollegeScorecard_Raw_Data_04192023.zip>

**FEATURE SELECTION AND IMPORTANCE ANALYSIS: GUIDING PREDICTIVE MODELING FOR GRADUATION RATES:**

The distribution analysis and exploratory data analysis (EDA) played a pivotal role in identifying features with strong predictive potential for graduation rates (C100\_4, C150\_4, and C200\_4) in four-year institutions.

A screenshot of a computer screen

Description automatically generated

Noteworthy features, mentioned below were selected based on their significant correlation with the target variables.

* FTFTPCTPELL (Pell Grant Rate)
* SAT\_AVG\_ALL (Average SAT Scores)
* TUITIONFEE1\_OUT (Out-of-State Tuition)
* DEBT\_MDN (Median Student Debt)
* UG25ABV (Non-Traditional Student Percentage)
* AVGFACSAL (Average Faculty Salary)
* UGDS (Undergraduate Enrollment)
* COST4 (Average 4-Year Program Cost)
* ADM\_RATE\_ALL (Overall Admission Rate)
* UGDS\_ASIAN (Percentage of Asian Students)

Correlational matrix analysis revealed that **SAT\_AVG\_ALL** exhibited a substantial correlation of 75% with all target variables, while **FTFTPCTPELL** displayed a negative correlation of approximately 70%. Additionally, **PLUS\_DEBT\_INST\_MD, AVG\_FAC\_SAL, and TUITIONFEE1\_OUT** demonstrated a correlation of 60% with all target variables.

Further insights from feature importance scores highlighted key drivers for understanding and predicting graduation rates. The top features include:

* FTFTPCTPELL (Feature Importance Score: 0.489644)
* SAT\_AVG\_ALL (Feature Importance Score: 0.134555)
* TUITIONFEE1\_OUT (Feature Importance Score: 0.050522)
* DEBT\_MDN (Feature Importance Score: 0.032774)
* UG25ABV (Feature Importance Score: 0.031669)

This insightful feature selection process guides the subsequent machine learning phase, emphasizing the pivotal role of financial aid, academic preparedness, and institutional characteristics in understanding, and predicting graduation rates.

**KEY OBSERVATION:**

1. **Demographics: Limited correlation, suggesting demographics may not strongly align with graduation rates.**
2. **Financial Aid Impact: High importance for FTFTPCTPELL and PLUS\_DEBT\_INST\_MD underscores financial aid's substantial influence.**
3. **Academic Preparedness: SAT\_AVG\_ALL and UG25ABV showcase academic preparedness as a significant factor.**
4. **Institutional Characteristics: AVGFACSAL and COST4 indicate the impact of institutional traits on graduation rates.**

**COMPARITIVE PERFORMANCE OF PREDICTIVE MODELS:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | TRAIN\_R2 | TEST\_R2 | TRAIN\_MSE | TEST\_MSE |
| LINEAR | 0.75 | 0.74 | 0.0077 | 0.0078 |
| POLYNOMIAL | 0.83 | 0.71 | 0.005 | 0.008 |
| POLYNOMIAL WITH LASSO | 0.78 | 0.77 | 0.006 | 0.007 |
| DECISION TREE | 0.87 | 0.63 | 0.003 | 0.011 |
| DECISION TREE TUNED | 0.83 | 0.81 | 0.005 | 0.005 |
| BAGGED LINEAR | 0.75 | 0.74 | 0.008 | 0.007 |
| BAGGED POLYNOMIAL | 0.80 | 0.76 | 0.008 | 0.008 |
| RANDOM FOREST | 0.81 | 0.76 | 0.005 | 0.007 |
| XGBOOST | 0.99 | 0.78 | 0.006 | 0.0002 |
| XGBOOST TUNED | 81 | 76 | 0.005 | 0.007 |
| RNN | 0.66 | 0.60 | 0.010 | 0.013 |
| Bidirectional RNN | 0.67 | 0.62 | 0.012 | 0.015 |
| RNN LSTM | 93 | 82 | 0.001 | 0.006 |
| RNN LSTM TUNED | 85 | 84.67 | 0.004 | 0.005 |

**EVALUATING MODEL PERFORMANCE: A CLOSER LOOK:**

**RNN LSTM TUNED**:

**INITIAL APPROACH WITH SIMPLE RNN**

My initial modeling efforts began with a simple RNN comprising three layers, which established a baseline for the model’s predictive capabilities. As I increased the number of neurons, I observed an improving trend in the training R2 scores. However, the test R2 scores eventually plateaued, indicating limitations to performance gains from increasing model complexity alone.

A graph with blue and orange lines

Description automatically generated

DEPTH EXPLORATION

I then fixed the neuron count to 128 and varied the number of layers to explore the effects of depth on the model. This exercise highlighted a nuanced relationship between model depth and performance, with a sweet spot beyond which additional layers did not translate into better generalization in unseen data.

A graph with lines and numbers

Description automatically generated

SEQUENCE LENGTH OPTIMIZATION

The pivotal moment in my analysis came when I altered the sequence length, which denotes the number of years of data used. A nine-year sequence, from 2012 to 2020, proved optimal, leading to the highest train and test R2 scores. This finding emphasized the importance of the temporal breadth of data in achieving predictive accuracy.

A graph with lines and numbers

Description automatically generated

BIDIRECTIONAL RNN AND OVERFITTING CHALLENGES

Subsequent experiments with a bidirectional RNN offered a marginal improvement but still indicated overfitting, as seen by the disparity between training and testing performance.

BREAKTHROUGH WITH LSTM TUNING

The real advancement came with the LSTM RNN model. With 9 features (d=9) across a 10-year sequence (T=10), I achieved an impressive train R2 of 85 and test R2 of 84.67, which demonstrated strong predictive performance without overfitting. This model configuration utilized an input size corresponding to the 9 features, a hidden layer size of 64 neurons, and 3 stacked LSTM layers, outputting predictions for the 4-year, 6-year, and 8-year graduation rates.

A red and green dotted diagram

Description automatically generated

ANALYTICAL INSIGHTS

In conclusion, the LSTM RNN's sophisticated gating mechanism proved to be highly adept at handling complex time-series data in the education sector. Its ability to remember and forget information selectively was key to capturing the nuances of educational pathways over a 10-year period.

**DECISION TREE REGRESSION**: **BALANCING FIT AND GENERALIZATION**

In my analysis, I utilized a Decision Tree Regression model to predict graduation rates. Initially, without hyperparameter tuning, the model showed signs of overfitting, as indicated by a high R2 score of 0.877 on the training set but a significantly lower score of 0.659 on the test set. This discrepancy was further evidenced by the difference in MSE, which was 0.0039 for the training set and 0.0107 for the test set.

A screenshot of a graph

Description automatically generated

To address the overfitting and enhance the model's generalization capability, I proceeded with hyperparameter tuning. The optimal parameters identified were {'max\_depth': 10, 'min\_samples\_split': 60}, with no restrictions on max\_features and max\_leaf\_nodes. This tuning process aimed to simplify the model, thereby reducing its complexity and potential to overfit.

A graph with orange and blue lines

Description automatically generated

After tuning, the model achieved an R2 score of 0.836 on the training set and a highly comparable 0.816 on the test set, demonstrating a commendable balance between fitting the training data and generalizing to unseen data. The MSE also reflected this improvement, with the testing set showing a modest increase to 0.0058, closely aligning with the training set's MSE of 0.0052.

A red and green dotted diagram

Description automatically generated with medium confidence

This Decision Tree model, particularly after tuning, offered better results than the untuned version due to its tailored complexity, which was sufficient to capture the underlying patterns in the data without being overly specific to the training set. The consistency between the training and testing performance metrics post-tuning underscored the effectiveness of the selected hyperparameters.

In conclusion, the Decision Tree Regression model's interpretability, and the improved generalization post-tuning position it as a valuable tool for stakeholders seeking to understand and act upon the factors influencing graduation rates.

**XGBOOST TUNED**:

This model achieved an R2 score of 0.8567 on the training dataset, indicating a strong fit to the training data. On the testing dataset, the R2 score was 0.7833, suggesting a slight decrease in predictive accuracy when generalized to unseen data, which could be indicative of mild overfitting.

The Mean Squared Error (MSE) was quite low, at 0.0045 for the training dataset and 0.0068 for the testing dataset, reinforcing the model's overall predictive power. This low MSE across both datasets emphasizes the model's ability to make accurate predictions with minimal error.

A graph of a graph of a graph

Description automatically generated with medium confidence

The hyperparameters played a crucial role in achieving this balance between fit and generalization. The final parameters were fine-tuned as follows: colsample\_bytree at 1, a learning\_rate of 0.1, max\_depth at 8, n\_estimators set to 1000, reg\_alpha (L1 regularization term) at 4, and subsample at 0.6.

Gradient boosting techniques like XGBOOST are adept at handling complex feature interactions through the boosting process, which builds models sequentially to correct the errors of the previous models. The careful tuning of hyperparameters like max\_depth and reg\_alpha helped to effectively manage the model's complexity, preventing it from becoming too specialized to the training data while still retaining strong predictive capabilities.

A red and green dotted diagram

Description automatically generated

In conclusion, the XGBOOST model stood out for its high R2 scores and low MSE, thanks to the strategic selection of hyperparameters that balanced model complexity with predictive accuracy, making it a robust choice for stakeholders looking for reliable graduation rate predictions.

**POLYNOMIAL WITH LASSO**:

During analysis, I applied Polynomial Ridge Regression to predict graduation rates, which yielded an R2 score of 0.837 on the training set. However, this model's performance on the testing set showed a significant drop to an R2 score of 0.713, suggesting a potential overfitting issue, as evidenced by the increased Mean Squared Error (MSE) from 0.0051 in the training set to 0.0090 in the testing set.

To mitigate this, I incorporated Lasso regularization into the Polynomial Regression, which not only penalizes the magnitude of the coefficients but also helps in feature selection by shrinking some coefficients to zero. By setting an alpha value of 50 and adjusting max iterations for convergence, I was able to obtain a more consistent model performance. The resulting R2 scores were 0.784 for the training set and a closely aligned 0.770 for the testing set. Similarly, the MSE showed a minimal difference, with 0.0068 for the training set and 0.0072 for the testing set.

A red and green dotted diagram

Description automatically generated

The near-identical R2 scores and MSE values across the training and testing datasets indicate a good balance between model bias and variance, a clear sign of the regularization effect. Lasso's ability to perform feature selection inherently through regularization resulted in a model that not only fits the data well but also generalizes better to unseen data, compared to the unregularized Polynomial Ridge Regression.

In conclusion, Polynomial Lasso Regression outperformed its unregularized counterpart by effectively addressing overfitting, as demonstrated by the more stable performance across both datasets. This model's ability to maintain a strong predictive performance without being too complex makes it a valuable tool for stakeholders needing reliable and interpretable predictions.

**RANDOM FOREST**:

In my exploration of predictive models for graduation rates, the Random Forest algorithm stood out due to its robust generalization capabilities. After an extensive grid search, which involved fitting 5 folds for each of 1008 candidates, totaling 5040 fits, I identified the best parameters for the model: {'max\_depth': 9, 'max\_features': 'sqrt', 'min\_samples\_leaf': 25, 'min\_samples\_split': 20, 'n\_estimators': 150}.

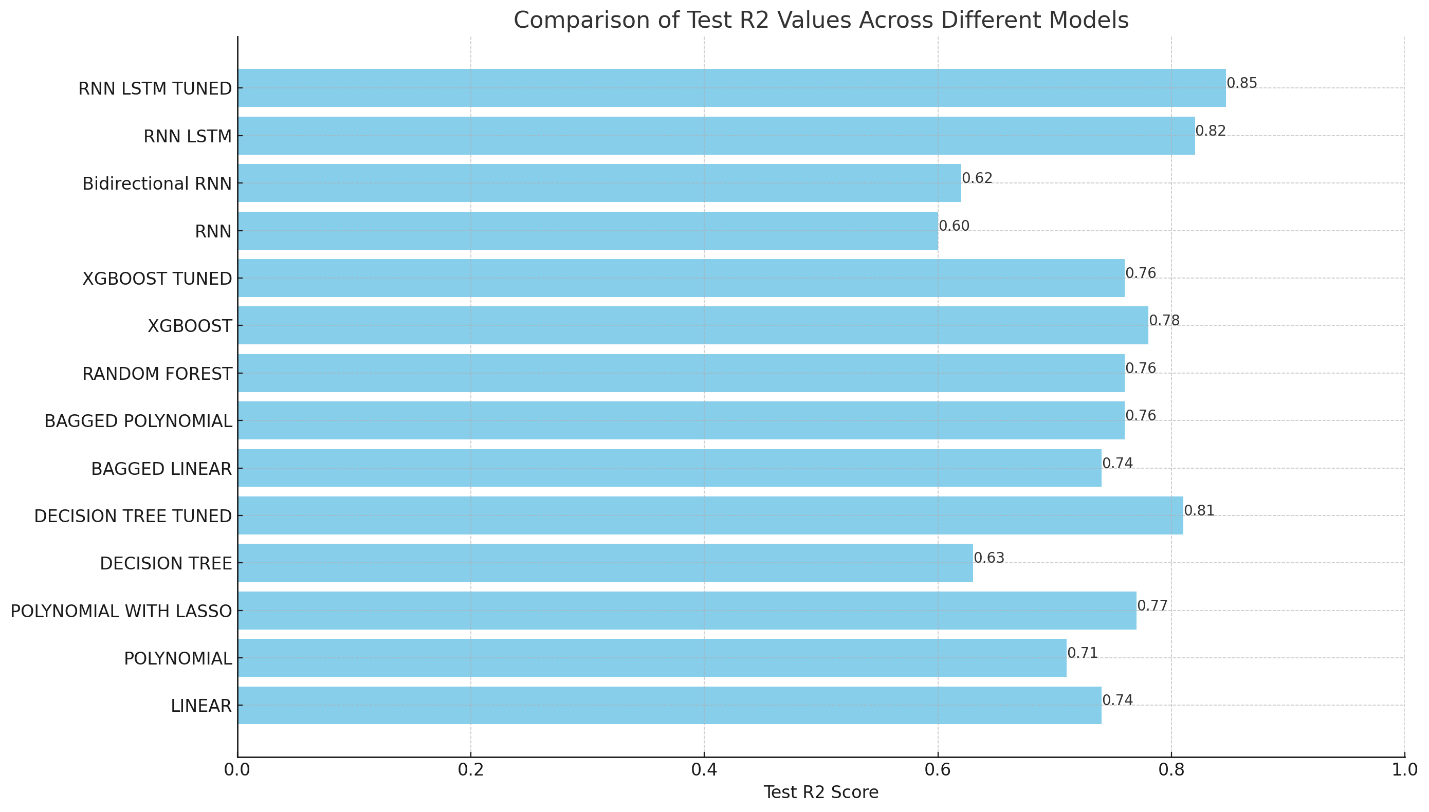
The tuned Random Forest model achieved an R2 score of 0.8115 on the training dataset and 0.7644 on the testing dataset, illustrating the model's ability to generalize well to unseen data. The Mean Squared Error (MSE) further attested to the model's consistency, with 0.0059 for the training set and 0.0074 for the testing set.

A red and green dotted diagram

Description automatically generated

The ensemble method employed by the Random Forest combines the predictions of multiple decision trees, which helps to improve generalization by reducing the variance that might be present in individual trees. The role of randomness in selecting features and samples to create diverse trees within the forest was crucial to prevent overfitting, which often plagues single decision trees.

**COMPARITIVE ANALYSIS OF MODEL TEST R2 SQUARES**



**TRAIN R2 SCORES AND TEST R2 SCORES**

A graph with blue and orange lines

Description automatically generated

Despite the RNN LSTM model's superior performance, with train and test R2 scores around 85%, indicating its robustness in predicting graduation rates, there remains room for enhancement.

**PERFORMANCE CEILING**: The current scores, while high, suggest that there is a performance ceiling that has not yet been breached. Further exploration into feature engineering, additional data sources, and advanced modeling techniques might reveal hidden patterns that could elevate the model's predictive accuracy.

**CONCLUSION**

* The RNN LSTM Tuned model excelled in predicting graduation rates, with robust train and test R2 scores near 85%.
* Key influences on graduation rates were highlighted, with financial aid and academic metrics like Pell Grant rates and SAT scores playing pivotal roles.
* Comparative analyses showed the XGBOOST model’s high training accuracy, but also its tendency toward overfitting, in contrast to the tuned Decision Tree's consistent generalization across datasets.
* Polynomial Regression with Lasso regularization addressed overfitting effectively, offering stable predictive performance.
* The ensemble approach of the Random Forest model underscored the strength of combining multiple decision trees to enhance generalizability and avoid overfitting.
* While the RNN LSTM model demonstrated impressive accuracy, opportunities for further improvements exist, potentially through the integration of advanced techniques like attention mechanisms, which could refine the model's focus on relevant patterns within the data.
* The findings equip education stakeholders with actionable insights for strategic planning to boost student success and institutional performance.
* Future work should explore broader datasets and more sophisticated model architectures, including the application of attention mechanisms, to potentially yield models with even greater predictive accuracy and insight.

**REFERENCES**

* Introduction to Machine Learning," taught by Professor Masoud Soroush at the University of Maryland Baltimore County (UMBC).
* <https://pytorch.org/tutorials/>