A Predictive Analytics Study:
Optimizing Recruitment Strategies for Teach For America

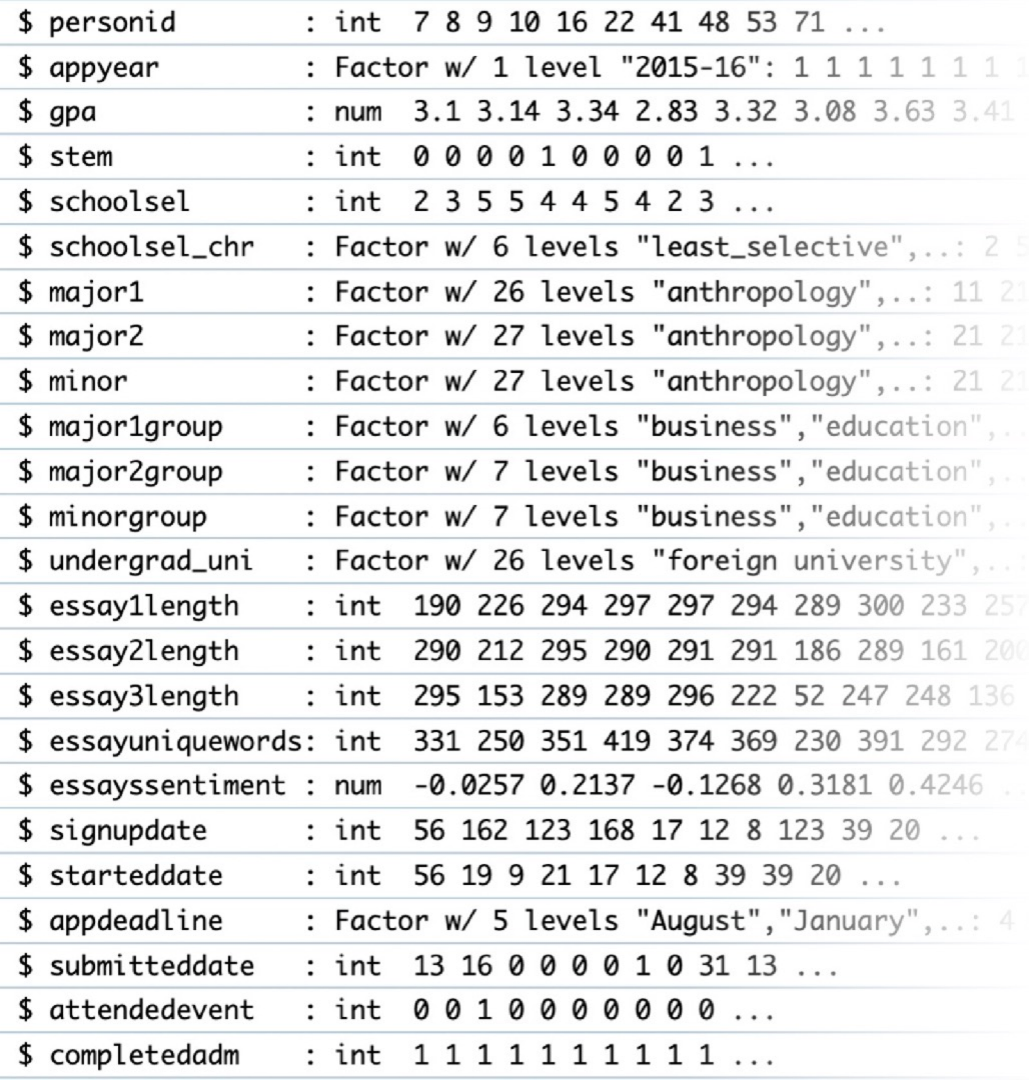
Developed by AdityaRaj Sharma

Introduction

- Teach For America is a nonprofit organisation with the mission to "enlist, develop, and mobilize as many of the nation's most promising future leaders to grow and strengthen the movement for educational equity and excellence.

- Challenge: The decrease in application rates directly impacts TFA's ability to maintain a robust teaching force, crucial for achieving its educational objectives in underserved communities.

- Solution: Leveraging advanced predictive analytics to provide TFA with strategic insights by analyzing applicant data. We aim to enhance TFA’s recruitment efficiency and retention strategies, ensuring a more impactful reach in communities needing quality education.



```
$ personid        : int  7 8 9 10 16 22 41 48 53 71 ...
$ appyear         : Factor w/ 1 level "2015-16": 1 1 1 1 1 1 1 1
$ gpa             : num  3.1 3.14 3.34 2.83 3.32 3.08 3.63 3.41
$ stem            : int  0 0 0 0 1 0 0 0 0 1 ...
$ schoolsel       : int  2 3 5 5 4 4 5 4 2 3 ...
$ schoolsel_chr   : Factor w/ 6 levels "least_selective",..: 2 5
$ major1          : Factor w/ 26 levels "anthropology",..: 11 21
$ major2          : Factor w/ 27 levels "anthropology",..: 21 21
$ minor           : Factor w/ 27 levels "anthropology",..: 21 21
$ major1group     : Factor w/ 6 levels "business","education",..
$ major2group     : Factor w/ 7 levels "business","education",..
$ minorgroup      : Factor w/ 7 levels "business","education",..
$ undergrad_uni   : Factor w/ 26 levels "foreign university",..:
$ essay1length    : int  190 226 294 297 297 294 289 300 233 257
$ essay2length    : int  290 212 295 290 291 291 186 289 161 200
$ essay3length    : int  295 153 289 289 296 222 52 247 248 136
$ essayuniquewords: int  331 250 351 419 374 369 230 391 292 274
$ essayssentiment : num  -0.0257 0.2137 -0.1268 0.3181 0.4246 ..
$ signupdate      : int  56 162 123 168 17 12 8 123 39 20 ...
$ starteddate     : int  56 19 9 21 17 12 8 39 39 20 ...
$ appdeadline     : Factor w/ 5 levels "August","January",..: 4
$ submitteddate   : int  13 16 0 0 0 0 1 0 31 13 ...
$ attendedevent   : int  0 0 1 0 0 0 0 0 0 0 ...
$ completedadm    : int  1 1 1 1 1 1 1 1 1 1 ...
```

The Dataset

Data Volume: Our analysis is based on a comprehensive dataset of 37,134 applicant records. This extensive data set allows for a nuanced understanding of applicant behaviors and traits.

Key Variables: The dataset encompasses diverse variables: GPA, STEM background, school selectivity, detailed academic majors and minors, groupings of majors/minors, undergraduate institution, quantitative and qualitative essay metrics, timeline metrics of applicant engagement with TFA, event participation, and the crucial outcome variable - completion of the admission process.

Objective: We aim to dissect these variables to predict the likelihood of an applicant completing the admission process, a key indicator for TFA's recruitment success.

Methodology

Tackling TFA's Recruitment Challenge

- Objective: Our primary goal is to develop a predictive model that accurately forecasts whether a candidate will accept a position with TFA. This forecast is pivotal in predicting the 'Completion of Admission Process' variable.

- Models Tested: We evaluated a spectrum of predictive models - kNN for its simplicity and effectiveness in smaller datasets, Naive Bayes for its efficiency with text data, Decision Trees for their interpretability, ANN for capturing complex, non-linear relationships, and SVM for its robustness in high-dimensional data spaces.

- Pre-processing for Models: After analyzing the dataset, we prepared our data for further analysis.
  - kNN – Converted categorical data into numeric data (factors) for analysis.
  - Naïve Bayes – Factorized categorical data
  - Decision Trees – Factorized categorical data
  - ANN – Converted categorical data into numeric data for analysis.
  - SVM – Converted categorical data into numeric data (factors) for analysis.

- Significance: Accurate predictions will empower TFA to allocate its resources efficiently, enhance recruitment and selection processes, and ultimately ensure a more profound impact in educational settings that most need skilled educators.

- Evaluation Criteria: Focus is on finding the model that best predicts a balanced outcome between admission completion and withdrawal by assessing the models ability of predicting the for criteria - True Positives, True Negatives, False Positive & False Negatives.

Analysis

K-Nearest Neighbour (kNN)

Reasons for Choosing this Model:

- Simplicity and effectiveness in smaller datasets.
- No assumptions about the data distribution.

- Model Setup: Utilizing a training and testing split of 80-20, we experimented with different 'k' values - 1, 3, 5, 7, and 11. This allowed us to assess the impact of neighbour consideration on the model's predictive accuracy.
- Performance Highlights: The model with k=5 showed optimal performance, striking a balance in predicting admission completion. It demonstrated high accuracy in identifying true positives.

Outcome:

Despite its effectiveness, we noticed a bias towards predicting positive outcomes, leading to a significant number of false positives. This bias could result in overestimating the number of successful admissions, potentially affecting TFA’s resource allocation and strategic planning.

k Value	True Negatives	True Positives	False Positives	False Negatives	Accuracy %
1	479	3939	593	361	82.24%
3	353	4135	719	165	83.54%
5	280	4218	792	82	83.73%
7	243	4252	829	48	83.67%
11	185	4283	887	17	83.17%

Analysis

Naïve Bayes

Reasons for Choosing this Model:

- Ideal for scenarios with categorical input variables.
- Efficient for large datasets and offers fast prediction times.
- Effective in text classification and when variables are conditionally independent.

Outcome:

The model was biased towards predicting that applicants would complete the admission process when they might not. This could lead to overestimation of successful admissions and may affect the overall efficiency of the admissions process.

Model Setup:

- Implemented an 80-20 split between training and testing datasets. The continuous variables were scaled for normalization.
- The GPA was categorized into bins (Low, Medium, High, Very High) to simplify its impact on predictions.

Performance Highlights:

- Demonstrated moderate accuracy with its strength in predicting true positives (applicants likely to complete the admission process). This is evidenced by the model's higher specificity in both the training and testing datasets.

Dataset	True Negatives	True Positives	False Positives	False Negatives	Accuracy
Training	885	15,576	3,403	1,624	76.61%
Holdout	220	3,897	852	403	76.64%

Analysis

Decision Trees

Reasons for Choosing this Model:

- Facilitates easy interpretation and understanding of the data.
- Efficiently handles both numerical and categorical data.

Outcome:

Across all configurations, a high number of false positives was observed, indicating a potential bias towards predicting completion of the admission process.

Model Setup: The dataset was split into training and test sets following an 80-20 ratio, and models with different trial numbers (10, 8, and 25) were evaluated to gauge the best trial.

Performance Highlights:

- Increasing the number of trials typically resulted in an increase in true negatives, suggesting a better capability to identify candidates unlikely to complete the admission process.
- The model with 8 and 25 trials offered a more balanced prediction between successful and unsuccessful admissions, proving effective in identifying both classes with good accuracy.

No. of Trials	True Negatives	True Positives	False Positives	False Negatives	Accuracy %
Original	70	4609	1106	75	79.84%
10	108	4563	1068	121	79.70%
8	108	4572	1068	112	79.86%
25	184	4480	992	204	79.59%

Analysis

Artificial Neural Network (ANN)

Reasons for Choosing this Model:

- Scalable to complex datasets and capable of learning intricate patterns.
- Flexibility in modelling through the adjustment of layers and nodes.

Model Setup: Configured an 80-20 split for training and testing. The model utilized a diverse set of input features including GPA, essay characteristics, and engagement metrics, with different numbers of hidden nodes tested for optimization. The model was trained and tested through hidden nodes.

Performance Highlights: The high accuracy of 80% may be primarily due to the model's ability to predict true positives, rather than its overall predictive performance.

Outcome:

The ANN model, particularly with 3 hidden nodes in its final iteration, showed an improved ability to correctly identify true negatives. However, the model still faced challenges in reducing false negatives.

Nodes	True Negatives	True Positives	False Positives	False Negatives	Accuracy
Original Model	2	4297	3	1070	80.03%
Hidden Nodes: 2	3	4294	6	1069	79.99%
Hidden Nodes: 3	4	4295	5	1068	80.03%

Analysis

Support Vector Machines (SVM)

Reasons for Choosing this Model:

- Particularly suitable for binary classification tasks with clear margin of separation.
- Versatile with different kernel functions to capture various data patterns.

Outcome:

- While the SVM model's training accuracy increases with higher cost values, the model's practical performance on the test data remains questionable.
- The increased training accuracy suggests that the model fits the training data better with a higher cost, but the testing phase highlighted the model's inability to detect the negative class effectively.

Model Setup:

- The data was modelled on a linear SVM first and then experimented with a Radial Basis Function (RBF) kernel to capture nonlinear patterns. Fine tuning was done by experimenting with different cost values

Performance Highlights:

- The linear kernel SVM produced a confusion matrix that revealed it failed to predict any negative class instances correctly.
- A slight improvement was observed with the RBF kernel but still missed many negative instances
- Higher cost values imply a higher penalty for misclassified points, thereby forcing the model to fit the training data more accurately.

Model Type	Kernel Type	True Negatives	True Positives	False Positives	False Negatives	Accuracy
Linear Model	Linear	0	4300	0	1072	80.04%
RBF Kernel	RBF	2	4299	1	1070	80.06%

Conclusion

TFA's Recruitment Process

Model	Performance	Bias	Practicality
kNN	Highest accuracy with k=5 (83.73%). Increase in true positives and decrease in true negatives as k increased.	Pronounced bias towards predicting positive outcomes.	Suitable for smaller datasets. Bias towards positive predictions could be problematic for balanced prediction needs.
Naive Bayes	Moderate accuracy, around 76.6% in both training and holdout datasets. High number of true positives.	Tendency to predict completions, leading to higher false positives.	Efficient in identifying likely completions, but struggles with balancing false positives and negatives.
Decision Trees	High accuracy with 8 trials (79.86%). Increase in true negatives and decrease in false negatives with more trials.	Less pronounced bias towards admissions compared to other models.	Balanced approach for identifying successful and unsuccessful admissions.
ANN	Consistent accuracy around 80%, but imbalance in sensitivity and specificity.	Struggles with negatives, indicating challenges in predicting non-completions.	Suitable for complex datasets but limited by convergence issues and bias towards completions.
SVM	Linear SVM: 80.04% accuracy, slight improvement with RBF kernel to 80.06%.	Heavy bias towards positive class, minimal ability to identify the negative class effectively.	Higher cost parameters improves training accuracy but doesn't translate to effective generalization.

Conclusion

TFA's Recruitment Process

Parameters	k-Nearest Neighbors (k=5)	Decision Trees (8 and 25 Trials)
Sensitivity to Imbalanced Data	Highly sensitive to class imbalances, affecting the accuracy and reliability of predictions.	Better equipped to handle class imbalances, offering more reliable predictions in diverse data scenarios.
Practicality for Complex Datasets	Performance may diminish with complex datasets containing a mix of categorical and numerical data.	More adept at handling complex datasets with various types of data, offering versatile applicability.
Scalability and Efficiency	Computationally intensive as dataset size grows, which might be challenging for TFA’s expanding data needs.	Generally more scalable and efficient in processing large datasets, suitable for TFA’s evolving data requirements.
Interpretability	Less Interpretable – difficult to trace back the paths of the predictive model.	Highly Interpretable - allows for easier understanding and communication of how predictions are made.

Recommendations

Decision Trees for Teach for America

- Balanced Accuracy: The model achieves a high level of accuracy while maintaining a balance between correctly predicting those who will complete and those who will withdraw from the admission process. This balance is essential for TFA's needs.

- Class Imbalance Management: Decision Trees effectively handle the class imbalance present in TFA's dataset. This is crucial for ensuring that predictions are not biased towards the more represented class.

- Interpretability: The model offers a high degree of interpretability, which is valuable for TFA's stakeholders. It allows for easier understanding and communication of how predictions are made, which is important for decision-making processes within TFA.

- Customization and Flexibility: Decision Trees can be easily customized and tuned to TFA's specific requirements. Adjustments can be made to cater to changes in the application process or applicant profiles over time.

- Simplicity and Efficiency: Compared to more complex models like kNN, ANN or SVM, Decision Trees are simpler to implement and require less computational power, making them efficient for TFA's operational context.

- Robustness to Overfitting: With the right tuning, particularly in the number of trials, Decision Trees can avoid overfitting, ensuring that the model generalizes well to new, unseen data.

- Insight Generation: Beyond prediction, this model can provide insights into the factors influencing an applicant's likelihood to complete the admission process, aiding TFA in strategic planning and policy formulation.

- Long-term Sustainability: The model's adaptability and straightforward nature make it sustainable for long-term use. As TFA's data and requirements evolve, the model can be updated and maintained without significant overhauls.

Thank You