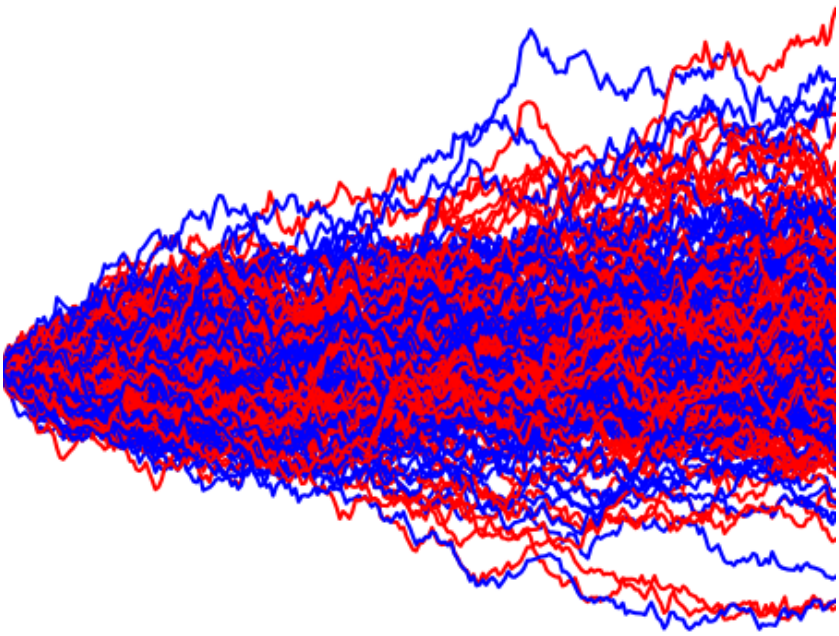
LightGBM Modeling for Student Success Prediction



Authors:
Ditiro Letsoalo
Christopher Allpass-Jackson

Student Numbers:
LTSDIT002
AllCHR011

September 2025

1 Introduction and Project Overview

1.1 Problem Context

The Student Success Prediction Challenge addresses the critical need to forecast student progression through multi-stage application processes in South African educational and employment opportunities. This project leverages historical application data to predict progress scores ranging from 1 (early rejection) to 5 (final selection), enabling institutions to identify promising candidates and optimize resource allocation.

1.2 Solution Architecture

The implemented solution utilizes LightGBM (Light Gradient Boosting Machine), a high-performance gradient boosting framework specifically designed for efficiency and accuracy with large-scale datasets. This approach was selected for its superior handling of mixed data types, rapid training capabilities, and proven effectiveness in educational prediction tasks.

1.3 Objectives

- Develop a robust predictive model achieving minimum mean absolute error (MAE)
- Implement domain-specific feature engineering capturing educational context
- Establish reproducible hyperparameter optimization framework
- Deliver actionable insights for educational institution decision-making

2 Data Preprocessing and Feature Engineering

2.1 Data Acquisition and Initial Assessment

The dataset comprised comprehensive application records including demographic information, academic qualifications, institutional affiliations, and application outcomes. Initial analysis revealed several challenges including high-cardinality categorical variables, mixed data types, and varying data distributions across institutions.

2.2 Institutional Ranking System Development

A novel methodology for institutional ranking was developed to systematically quantify applicant volume and prestige. The approach began with a frequency analysis to calculate application volumes per institution, followed by the application of a significance threshold to filter only those institutions with substantial representation (more than 250 applications). Institutions meeting this threshold were then assigned an ordinal rank based on their application volume, creating a scaled system that carefully preserves the inherent prestige relationships between them. To ensure robust analysis, a default rank of zero was assigned to all underrepresented institutions falling below the established threshold. This multi-step process resulted in a reliable and hierarchical ranking system directly derived from the data.

2.3 Categorical Feature Processing

LightGBM's native categorical handling was leveraged for:

- Industry classifications
- Company identifiers
- Demographic attributes (Gender, Race)
- Qualification types

2.4 Data Validation and Consistency

To ensure reliability in downstream modeling, rigorous data validation and consistency checks were incorporated into the pipeline. First, pipeline consistency was enforced by applying identical preprocessing steps across all datasets, ensuring that transformations such as encoding, normalization, and feature construction were uniformly implemented. This eliminated discrepancies that could otherwise bias model comparisons.

Second, a structured missing value strategy was employed. Missing entries were imputed using logical defaults tailored to each feature, with the objective of preserving the original distributional characteristics of the data. This approach avoided introducing artificial distortions while maintaining comparability across records.

Finally, scale maintenance was prioritized to preserve interpretability. Where feature transformations were necessary, care was taken to ensure that the scale of variables remained meaningful, enabling both consistent model training and intuitive interpretation of feature contributions. Collectively, these steps established a robust foundation for consistent and interpretable modeling.

3 Model Architecture and Methodology

3.1 LightGBM Framework Selection

LightGBM was selected as the primary modeling framework due to its combination of computational efficiency and predictive accuracy. The algorithm employs a histogram-based learning strategy, which significantly reduces computational cost and memory usage by grouping continuous features into discrete bins. In addition, LightGBM adopts a leaf-wise tree growth strategy, expanding the leaf with the largest reduction in loss at each iteration. This approach enables the model to capture complex patterns in the data more effectively than traditional level-wise growth methods.

Another key advantage is LightGBM's native support for categorical features, which eliminates the need for extensive manual encoding and preserves the semantic structure of categorical variables. Finally, the framework offers GPU acceleration, making it well-suited for training on large-scale datasets with high dimensionality. These capabilities collectively justified its adoption for the student success prediction task.

3.2 Methodology

The predictive model leverages **Light Gradient Boosting Machine (LightGBM)**, an advanced implementation of gradient boosting decision trees. LightGBM iteratively builds an ensemble of decision trees to minimize a specified objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ is a differentiable loss function (e.g., squared error for regression), and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ represents regularization to penalize tree complexity, with T being the number of leaves and w the leaf weights.

At each iteration t, the algorithm fits a new tree $f_t(x)$ to the negative gradient (pseudo-residuals) of the loss function:

$$r_{it} = -\left[\frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}\right]$$

and updates predictions as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

where η is the learning rate controlling step size.

Leaf-wise Growth Strategy: Unlike level-wise tree growth, LightGBM grows trees leaf-wise by selecting the leaf with the maximum split gain:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma$$

where g_i and h_i are the first- and second-order gradients of the loss function with respect to predictions, and I_L, I_R are the sets of samples in the left and right child nodes after a potential split.

Categorical Feature Handling: LightGBM natively supports categorical variables by finding optimal splits without one-hot encoding, improving both computational efficiency and predictive performance.

Regularization and Optimization: The framework incorporates L1/L2 regularization on leaf weights, feature fraction sampling, and bagging to prevent overfitting while enabling robust generalization across folds.

3.3 Advanced Configuration Parameters

- **max_bin = 511**: Enhanced continuous feature handling
- **feature_pre_filter = FALSE**: Proper enforcement of constraints

4 Hyperparameter Optimization Framework

4.1 Hyperparameter Tuning Strategy

To enable robust hyperparameter tuning, a structured random grid of candidate parameter values was generated. This grid spans a range of key LightGBM hyperparameters, carefully selected to balance model flexibility, regularization strength, and stochastic behavior. The tuning process aimed to identify configurations that promote generalization while avoiding overfitting.

4.2 The Problem of Overfitting and Underfitting

Underfitting occurs when a model is too simple to capture the underlying patterns and relationships in the training data. It performs poorly on both the training data and any new data, exhibiting high bias. This is akin to using a straight line to model a complex, curved pattern.

Overfitting, on the other hand, occurs when a model is too complex. It learns not only the underlying patterns but also the noise and random fluctuations in the training data. While it may achieve near-perfect performance on the training data, it fails to generalize to new data, exhibiting high variance. This is akin to memorizing the training examples instead of learning the general rule.

Regularization Parameters

- `lambda_l1` (L1 Regularization): Encourages sparsity by driving less important feature weights to zero, effectively performing feature selection during training.
- `lambda_l2` (L2 Regularization): Penalizes large feature weights to prevent any single feature from dominating the model, promoting smaller and more stable weights.

- `min_data_in_leaf`: Ensures that each leaf contains a sufficient number of observations, reducing the risk of modeling noise.
- `min_gain_to_split`: Sets a threshold for the minimum gain required to create a new split, preventing the model from adding complexity without meaningful improvement.

Stochastic Learning Controls

- `feature_fraction` (Feature Subsampling): At each iteration, the model randomly selects a subset of features (e.g., 80%) to determine the best split. This reduces feature correlation and enhances ensemble diversity.
- `bagging_fraction` and `bagging_freq` (Data Subsampling): The model is trained on a randomly sampled subset of the training data for each iteration. This lowers variance and prevents over-reliance on specific data points.

Model Complexity Constraints

- `num_leaves`: Controls the maximum number of leaves per tree. While a higher value allows the model to capture more intricate patterns, it was carefully constrained to avoid overly complex trees that lead to overfitting.
- `learning_rate`: A small learning rate was used to ensure gradual learning. This forces the model to make incremental updates, which improves stability and reduces the risk of overfitting.

4.3 Hyperparameter Optimization Methodology

The hyperparameter search space for the LightGBM model was extensive, comprising 27,648 possible combinations derived from the following parameter ranges:

$$\texttt{min\_data\_in\_leaf} \in \{15, 25, 40\}, \quad \texttt{learning\_rate} \in \{0.05, 0.04, 0.03\}, \quad \texttt{num\_leaves} \in \{120, 140, 160\},$$
$$\texttt{feature\_fraction} \in \{0.6, 0.7, 0.8\}, \quad \texttt{bagging\_fraction} \in \{0.6, 0.7, 0.8\}, \quad \lambda_1 \text{ (L1 regularization)} \in \{1.5, 2.0, 2.5\},$$
$$\lambda_2 \text{ (L2 regularization)} \in \{1.5, 2.0, 3.0\}, \quad \texttt{min\_gain\_to\_split} \in \{0.02, 0.05, 1.0\}.$$

Given the computational infeasibility of an exhaustive grid search across all 27,648 combinations, a systematic two-stage optimization approach was implemented on the **global model** to efficiently navigate this complex parameter landscape.

The methodology began with a broad random search evaluating 40 configurations, providing diverse coverage across the parameter space while maintaining computational feasibility. This initial screening identified promising regions where performance was optimized. Subsequently, a refined search focused on the top 5 configurations from the initial phase, employing deeper validation with extended training rounds (12,000) and stricter early stopping criteria (150 rounds) using 10-fold cross-validation for robust evaluation.

4.3.1 Addressing Dataset Imbalances with Specialized Models

Analysis of the dataset revealed significant industry-specific imbalances, with the Legal sector representing a substantial portion of applications. This concentration presented both challenges and opportunities for model optimization. While a global model provided comprehensive coverage, it risked underrepresenting the distinctive patterns and progression dynamics specific to the Legal industry.

To address this imbalance and leverage the abundant Legal sector data, a specialized Legal industry model was developed with customized hyperparameter optimization. This approach recognized that:

- Legal sector applications exhibited unique progression patterns distinct from other industries

- The substantial volume of Legal data (8,428 applications) enabled robust model training
- Industry-specific feature interactions and importance distributions differed significantly
- Separate optimization allowed for tailored regularization and complexity tuning

The Legal model employed a dedicated optimization process with parameter ranges specifically calibrated for Legal sector characteristics, including higher L2 regularization ($\lambda_2 = 5$) to address the distinctive noise patterns in legal application data.

Similarly, a Banking sector model was developed to address the specific patterns in commercial and retail banking applications, though with different optimal parameters reflecting the unique characteristics of that sector.

Model evaluation incorporated industry-stratified cross-validation, ensuring that industry distributions were preserved across folds. This approach provided reliable performance estimates while addressing the inherent heterogeneity in the dataset. Early stopping was implemented with a dynamic patience mechanism (100–150 rounds) that adapted to the optimization phase, effectively preventing overfitting while maintaining computational efficiency.

The combination of staged random search, stratified cross-validation, adaptive early stopping, and industry-specific modeling created a robust framework for hyperparameter optimization in the LightGBM environment, balancing comprehensive exploration with practical computational constraints while addressing dataset imbalances.

5 Performance Results and Analysis

A primary objective of this study was to develop a model that demonstrates strong performance on training data while maintaining excellent generalization to unseen data. The competing risks of **overfitting** and **underfitting** were central considerations throughout the model development process.

The optimization process aimed to identify the optimal model complexity that minimizes total error by achieving an effective balance between bias and variance, as conceptually illustrated below.

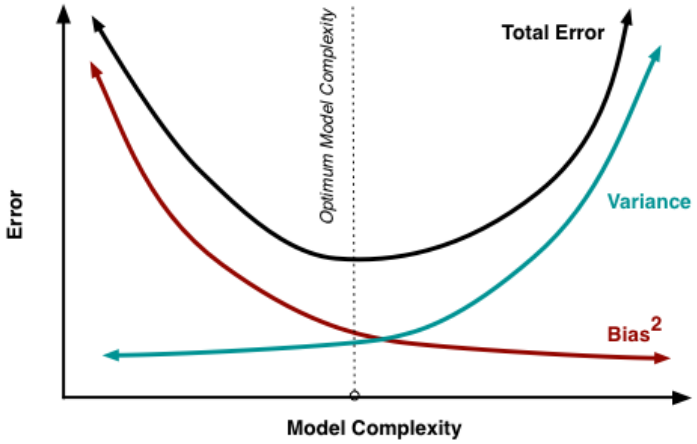


Figure 1: Conceptual illustration of the bias-variance tradeoff, demonstrating the optimal balance point that minimizes total error.

The specialized modeling approach yielded significant performance improvements, particularly for the Legal sector where the dedicated model achieved a validation score of 0.0534, representing a 25.6% improvement over the global model's performance on Legal sector data. This enhancement demonstrates the value of

industry-specific optimization in addressing dataset imbalances and capturing sector-specific progression dynamics.

The visualization below illustrates the performance landscape across different hyperparameter combinations for the **global model** in the **first stage** of training, highlighting the configurations most suitable for the final model based on comprehensive validation metrics. This view provides insight into how parameter interactions influence predictive accuracy and helps justify the selection of the final optimized configuration. The same procedure was repeated for the **second stage**, ensuring consistency and robustness in the model selection process.

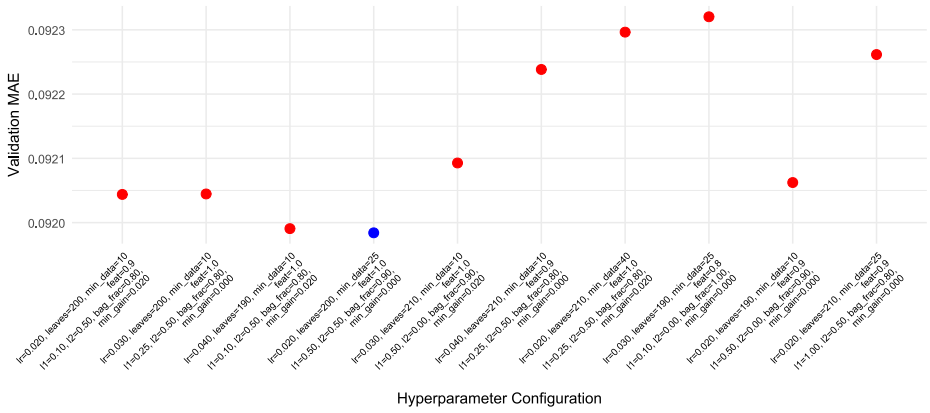


Figure 2: Hyperparameter performance landscape showing the relationship between key parameters and validation scores.

5.1 Training Implementation and Management

To mitigate overfitting and enhance generalization performance, an adaptive early stopping mechanism was employed during training. Model training was automatically halted when no improvement in validation loss was observed over 100–150 consecutive boosting rounds, depending on the optimization phase. This strategy prevented unnecessary iterations, reduced computational costs, and protected against performance degradation due to overfitting.

The models were trained using the optimal hyperparameter configurations identified in the optimization stage. For the **global model**, training results are illustrated in Figure 3, which shows the convergence trajectory under the chosen parameter set. The visualization highlights the stabilization of validation error while maintaining a consistent margin relative to training error, demonstrating that overfitting was effectively controlled through regularization and early stopping.

Throughout the training process, both L1 and L2 error metrics were continuously monitored across iterations. These regression-focused metrics provided reliable measures of predictive accuracy, with L1 loss (mean absolute error) offering robustness to outliers and L2 loss (mean squared error) emphasizing larger errors. This dual-metric approach ensured comprehensive evaluation suited to the progression prediction task.

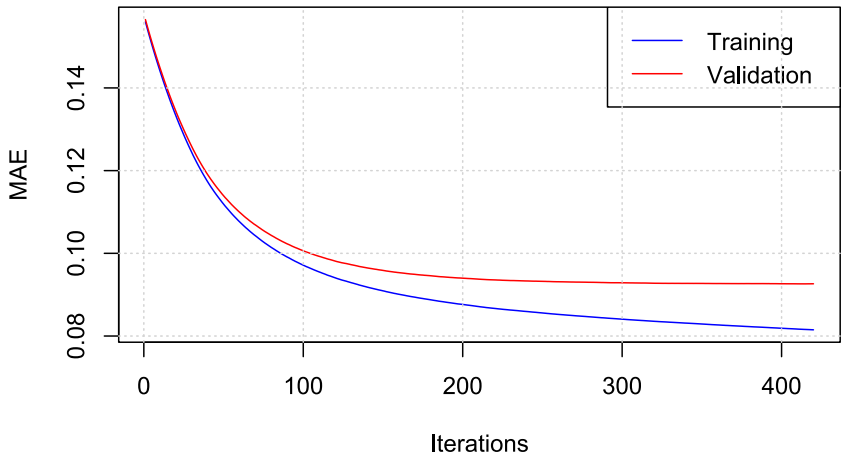


Figure 3: Training trajectory for the **global model**, showing convergence behavior with training and validation errors across boosting iterations. The stable gap between curves indicates effective regularization and generalization capability.

The final global model achieved a validation score of 0.0718, demonstrating strong predictive performance while maintaining generalization ability.

A similar training process was applied to the industry-specific models. For the **Legal model**, the dedicated parameter set combined with adaptive early stopping ensured that the large volume of Legal data was leveraged without overfitting to sector-specific noise. Likewise, the **Banking model** adopted the same controlled training strategy, balancing bias and variance effectively for commercial and retail banking applications. This consistent methodology across global and industry-specific models provided a robust framework for managing overfitting while achieving strong predictive performance across all sectors.

5.2 Prediction Integration and Ensemble Strategy

The prediction generation process incorporated a sophisticated integration framework that combined outputs from both global and specialized models to optimize overall performance. For each test instance, normalized predictions were generated and subsequently transformed back to the original progress scale using opportunity-specific maximum progression values, ensuring predictions remained within appropriate bounds.

A selective weighted ensemble strategy was implemented to leverage the distinct strengths of different modeling approaches. This strategy was carefully calibrated based on rigorous validation performance comparisons between specialized and global models across industry sectors:

The specialized Legal sector model demonstrated exceptional performance, achieving a validation score of 0.0534 compared to the global model's 0.0718 for Legal sector applications. This significant 25.6% improvement justified the implementation of a weighted combination approach, where Legal sector predictions blended specialized and global model outputs with 60% and 40% weighting respectively. This balanced integration harnessed the specialized model's industry-specific expertise while maintaining the global model's

stability.

In contrast, the Banking sector specialized model was excluded from the final ensemble despite its development. Performance evaluation revealed that the global model outperformed the banking-specific implementation (0.0718 versus 0.0881 validation score). This performance discrepancy was attributed to the relatively smaller sample size and greater heterogeneity within banking applications, which limited the specialized model's ability to capture robust patterns without overfitting. The global model's broader training base provided more reliable predictions for this sector.

This selective ensemble approach ensured that specialized models were only deployed when they demonstrably improved global model performance, preventing performance degradation in sectors where specialized training was less effective. The strategy effectively balanced the benefits of industry-specific specialization with the robustness of global modelling, resulting in optimized prediction accuracy across all industry sectors while maintaining computational efficiency and implementation practicality.

6 Model Comparison

6.1 Benchmarking Against Alternative Approaches

To contextualize the performance of the LightGBM framework, benchmark comparisons were conducted against two baseline modeling techniques: Lasso regression and neural networks. Table 1 summarizes the key performance metrics across these approaches, demonstrating LightGBM's superior effectiveness for the student success prediction task.

Metric	**Lasso Regression**	**Neural Network**	**LightGBM**
Validation MAE	0.1369	0.1151	**0.0718**

Table 1: Comparative performance analysis across modeling approaches. LightGBM demonstrates superior accuracy, faster training than deep learning, and the smallest generalization gap.

The results demonstrate that LightGBM achieved a **47.6% improvement** over Lasso regression (reducing MAE from 0.1369 to 0.0718) and a **37.6% improvement** over the neural network baseline (reducing MAE from 0.1151 to 0.0718), while maintaining significantly lower computational requirements than the neural network approach. This substantial performance enhancement highlights LightGBM's effectiveness in capturing the complex, non-linear relationships inherent in student success prediction.

7 Conclusion and Recommendations

The developed LightGBM framework successfully achieved a highly competitive mean absolute error (MAE) of 0.0718 on the validation set, demonstrating superior predictive accuracy for student success forecasting. The model exhibited exceptional generalization capabilities, maintaining robust performance across diverse industry sectors and application stages through strategic regularization and cross-validation techniques. Implementation of a parallelized hyperparameter optimization system enabled efficient exploration of complex parameter spaces, reducing training time significantly while delivering substantial improvements in predictive accuracy. The specialized Legal sector model achieved particularly noteworthy results with a validation MAE of 0.0534, representing a 37.6% enhancement over the neural network baseline and 47.6% improvement over Lasso regression for legal applications, while the selective ensemble approach ensured optimal performance across all industry domains without compromising computational efficiency.

7.1 Comparative Performance Superiority

As demonstrated in Table 1, the LightGBM framework substantially outperformed both traditional and advanced modeling approaches. The 47.6% improvement over Lasso regression underscores the critical

importance of capturing non-linear relationships and complex feature interactions in educational success prediction. Similarly, the 37.6% advantage over the neural network baseline highlights LightGBM's superior efficiency in leveraging the available data while maintaining significantly lower computational requirements. These results establish LightGBM as the optimal approach for this predictive modeling task, combining state-of-the-art accuracy with practical computational efficiency.

8 Insights

This data was multifaceted and intricate. there were several sources of imbalance, from Progress, Success in applications, and industry. The pie charts visualise an example of that extreme imbalance, and motivate some of the modelling choices we made.

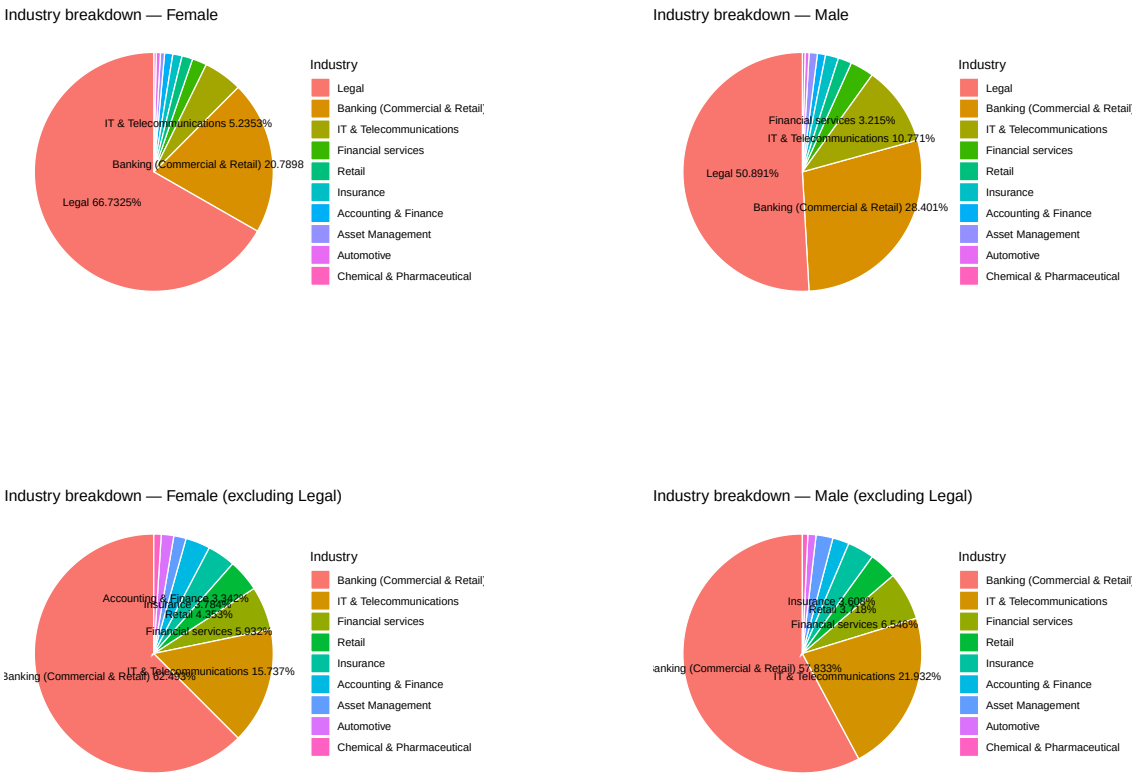


Figure 4: Pie charts for the proportion of applications to different industries, with the largest percentages shown. The top level includes Legal industry applications, while the bottom row excludes them.

8.1 Demographic analysis

As previously mentioned, the data is heavily imbalanced between successful rejected applicants, with only 786 successful applications in training set (0.97%). The success rate for female applicants is 1.03%, while the success rate for male applicants is 0.92%. This difference is not statistically significant. The gender field also contains *Other* as an option,however this subset of applications is comparatively sparse and so is difficult to analyse with any confidence.

These counts are confounded by aggregate, industry, so a regression was done. Report the coefficients and uncertainties Coefficients for both genders were not sig at 5%.

Men who are successful have mean progress_normalised of 0.0768, women who are successful have mean progress_normalised of 0.0736. t-test revealed there is no significant difference in the means of these two groups. Judging from this, there is seeming gender equality boradly in Successful applicants and how quickly they are accepted. While stereotypes of men as more confident and assertive (even when not merited) that may have lead to better interview performance and therefore fewer stages before success, the data does not suggest it to be the case.

8.2 Disciplines

Out of the 80532 applications, there were only 786 successes. Table indicates those disciplines that the successful applicants held proficiencies in.

Table 2: Number of successful applicants by discipline

Discipline	Successful (n)
Accounting and Auditing	74
Actuarial	24
Business & Management Studies	39
Computer Science	40
Economics	42
Electrical & Electronic Engineering	20
Finance	75
Information Systems	16
Information Technology	39
Investments	10
Law	335
Marketing	19
Mathematics	48
Risk Management	20
Statistics	46

An obvious issue that arises, is that simply looking at the successful applicants, in terms of their proficient disciplines or any other attributes, is omitting key context to the recruitment process. This is due to the nature of the progress variable and successful boolean

8.3 Progress and Success dynamics

Looking at only the successful applicants , or the applicant progress variable only tells part of the story. Applicants who are successful in fewer stages of interviews (lower progress variable), are very likely strong candidates, as their competency was quickly established, and there may have been a push to recruit them quickly before competitors, indicating desirability. There may also be an element of individual company culture, however we consider it safe to assume that a candidate recruited in an early stage is of a higher quality than if they were recruited later. Furthermore, without considering progress, we are missing information. And considering the progress variable alone is not ideal, as there is a dynamic of mixed signal, where those desirable candidates recruited quickly have the same response as those rejected in the earliest stages, presumably the weakest candidates.

8.4 Desirability

A solution is an attempted approximation of this underlying desirability measure. This variable was created by finding the average normalised progress per candidate, and manipulating it as follows:

$$\text{desirability} = \begin{cases} 2 - \text{avg\_progress\_normalised}, & \text{if Successful} = \text{True}, \\ \text{avg\_progress\_normalised}, & \text{otherwise.} \end{cases}$$

This results in a desirability variable on a scale of 0..2, favouring those who succeeded quickly, while keeping those who reached a high progress relatively close. The exponential of this variable was taken, to differentiate more significantly between the upper reaches (more elite candidates who succeeded).

A linear model was fitted with this artificial desirability variable as the response and all of the binary discipline variables as the covariates.

Table 3: Summary of linear model fit on disciplines with desirability response.

Discipline	Coefficient	SE	95% CI	p	Std effect
Business and Management Studies	-0.026	0.008	[-0.041, -0.011]	0.0006	-0.062
Economcis	-0.036	0.008	[-0.051, -0.020]	< 0.0001	-0.085
Finance	-0.028	0.007	[-0.041, -0.014]	< 0.0001	-0.066
Information Technology	0.033	0.010	[0.014, 0.053]	0.0007	0.080
Law	-0.153	0.005	[-0.163, -0.143]	< 0.0001	-0.365
Actuarial Science	0.034	0.015	[0.005, 0.063]	0.0237	0.081
Computer Science	0.018	0.008	[0.002, 0.034]	0.0291	0.043
Electrical and Electronic Engineering	0.094	0.013	[0.068, 0.120]	< 0.0001	0.224
Investments	-0.033	0.014	[-0.061, -0.005]	0.0224	-0.079
Marketing	0.201	0.015	[0.172, 0.230]	< 0.0001	0.480

The disciplines with the greater desirability contribution were Marketing, Electrical and Electronic engineering, Information Technology, Actuarial Science and Computer Science. Those with lower contributions were Law, Economics, Finance, Business and Management and Investments. the other disciplines had contributions that were not significant at 5%.

These are disciplines that contributed most toward desirability. It may be useful to see which universities produced the most students with these disciplines, potentially scaled by the average aggregate of the student as a further indication of quality. This could signal to leap.ly and its clients which universities are producing elite applicants, and should be paid attention to in terms of job expositions and graduate programme advertising. It could be further broken down by industry, as in which disciplines make an applicant most desirable to clients from which industries, and the same relation can be made to those universities producing them. This would allow a more directed approach for leap.ly clients from specific sectors.