I think I'll go to law school today! Gender and Racial Parity in the Academic Legal Sphere

Malthe Rødsgaard Pabst Lauridsen (mrla@itu.dk), Nicola Clark (niccl@itu.dk)
IT University of Copenhagen, May 20, 2024

1 Abstract

This paper investigates how geometric impacts the fairness of a random forest model and a logistic regression model, trained to predict if a new law student will eventually pass the bar exam. AI algorithms are a tool now commonly used as part of many universities' admissions processes, and as such, the fairness of such models can have a large impact on who become the next generation of experts. De-biasing the data was found to reduce accuracy for both the random forest model and the logistic regression, but increased fairness of the random forest significantly. For the de-biased models the importance of the *LSAT* score feature decreased and the importance of the *fulltime* study feature increased for the Random forest model. Bias university admission models can have a huge impact on an individual's life and broader society. De-biasing data prior to training and utilising explainability methods should therefore be standard to avoid further perpetuating bias. This papers findings indicate that care should be taken to explore the remaining bias in the data, as methods used to combat bias have varying degrees of effectiveness depending on the data and the model.

2 Introduction

This paper aims to explore and investigate fairness in the context of the law school bar exam dataset from the LSAC National Longitudinal Bar Passage Study [1].

The typical and most common pathway to becoming a qualified lawyer in America involves going to college, taking the LSAT, attending law school, and finally, passing the bar exam. Law school admissions in America center heavily upon LSAT performance and perceived potential to academically thrive and go on to pass the bar. In recent times machine learning has played a growing role in the process used to sort and select hopeful university applicants [2]. Such methods introduce numerous social and ethical conundrums around the topic of the fairness. Are the decisions made by these models, which judge the suitability of candidates fair, and what metrics can be used to explore this. Furthermore, are tests, such as the bar exam, fair themselves? Or are there features that predict the outcome of the bar exam more than what the test is aiming to measure: an individuals suitability and competence as a potential future lawyer? Features such as parental income for example, may relate to passing the bar, but they may also correlate with the protected feature of race.

This paper utilises the law school dataset to compare and contrast bias and values for a fair-

ness metric, for two classifiers trained on the data, a random forest model and a logistic regression model. These models were set up to mimic the types of algorithms that may be used by law school selection committees. Furthermore, this paper investigates if training these models using a geometric re-projection of the data has an impact on the fairness and performance of the models. Additionally, the explainability of the models is reflected upon by examining global feature importance's of both models, alongside considering the impact on the broader societal issues of fairness, ethics, and philosophy.

3 The Dataset

The dataset is from the LSAC National Longitudinal Bar Passage Study [1] from 1999. It was originally collected to investigate anecdotal reports of low bar pass rates amongst individuals who where not *White*. The dataset contained information about the pass rates of the bar exam, alongside demographic information and data on academic attainment such as LSAT test scores. A full description of the raw dataset can be found in appendix A. The dataset's target variable was whether or not a student passed the bar exam on their first attempt, furthermore, there were 2 protected attributes: sex and race. The racial categories in the data were: *White*, *Black*, *Asian*, *Hispanic*, and *Other*, while the sex categories were *Male* and *Female*.

In order to clean the data, many columns, such as *ugpa*, were dropped as they correlated highly with, or were identical to, other features. Other columns, such as *decile1b*, were dropped due to lack of documentation and uncertainty as to their precise content. An average ranking column was created from the remaining first and third year university ranking columns *decile1* and *decile3*, as these features were highly correlated with one another.

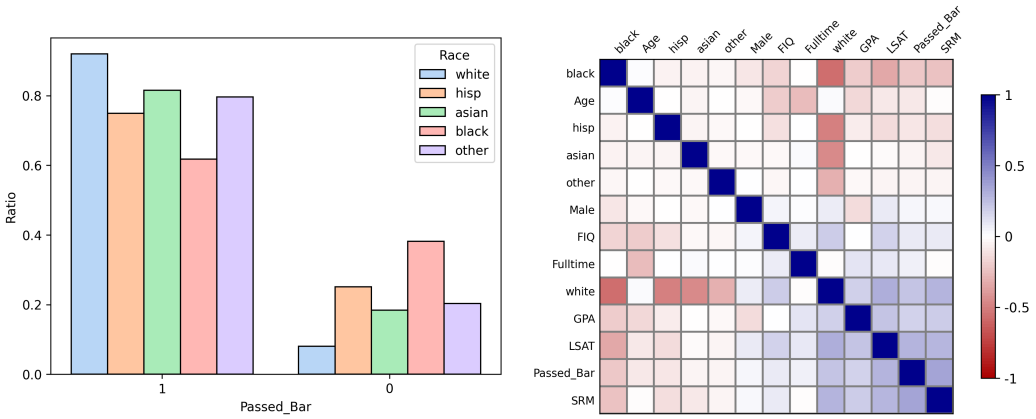
3.1 EDA

The cleaned dataset consisted of 20466 rows and 13 columns. Six of these columns related to the protected attributes of sex or race and one was the target variable, this left six features on which later models could be trained. These features were an individuals age, their family income quantile (FIQ), whether or not they were studying fulltime, their LSAT score, and their schools' ranking.

The data consisted of 56.3% males and 43.7% females, and the race distribution was predominantly *White* with 83.9%, 5.9% was *Black*, 4.5% *Hispanic*, 3.9% *Asian*, 1.8% *Other*. When inspecting the target variable, it was found that 11.2% of individuals had failed the bar, while 88% had passed. This was further inspected by investigating the pass rate for each of the categories within each protected variable. Figure 1(a) shows that the pass rate for *White* individuals was substantially higher than that of other races, with the pass rate of *Black* people being the lowest; 10% lower than that of

the next lowest racial group.

A correlation heatmap was created to explore the relationship between the features (Figure: 1(b)). From this, one can see that the *White* demographic positively correlated with features such as FIQ, LSAT, and School Ranking. This was not the case for other racial groups, especially *Black* individuals, where a negative correlation occurred for the same features. Aside from the racial features, positive correlations exited between most other features and the target variable, the strongest correlation being the schools' ranking.



(a) Bar pass and failure rates by race

(b) Correlation heatmap of all features

Figure 1: A barplot showing the ration of pass and failure rates for each race (a) and a correlation heatmap of all the features (b).

The distributions of those that passed compared to those that failed the bar were investigated by grouping by race. Figure 2 shows that FIQ was lower for those that failed the bar. This was especially the case for those in the group *Other*. Interestingly, this pattern does not hold for *Black* people in the dataset where there is no group-wide discrepancy in income levels depending on passing or failing the bar. No large differences were found between groups when the same procedure was followed with a grouping by sex.

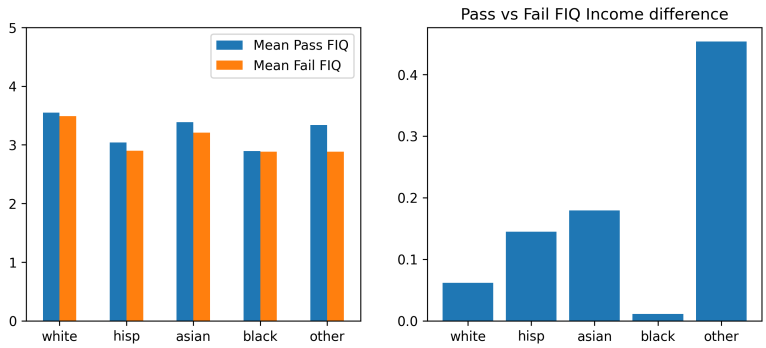


Figure 2: FIQ difference when passing vs failing the bar exam by race

4 Methods Explained

The following section will briefly explain some of the methodologies utilised to investigate the LSAC dataset.

Geometric re-projection

Geometric re-projection is a way of making features less biased by removing linear correlations between the protected attributes and the features from the data[3]. This method involves projecting standardised features onto a subspace where $Corr(x_j, r_j)$ is maximised under the constraint $Corr(r_j, p_i) = 0$. Where x_j are the original feature vectors, r_j are the projected feature vectors, and p_i are the protected feature vectors. This subspace is spanned by the orthonormal basis of the protected features p_i, such that:

$$r_j \cdot p_i = 0$$

A parameter λ can be used to regulate the level of fairness of the projection. Where r'_j is the partially projected data with fairness level λ.

$$r'_j = r_j + \lambda(x_j - r_j)$$

When $\lambda = 0$ then $r'_j = r_j$ and the data is de-biased as r'_j is orthogonal to the protected features; there are no linear correlations. When $\lambda = 1$ then $r'_j = x_j$ and no linear correlations are removed.

Fairness Metrics

There exist many different metrics to measure fairness. However some of the more broadly utilised ones are statistical parity, equalised outcomes, and equalised odds.

Statistical parity is when the rate of selection amongst all groups is equal. For example $P(Selected|A) = P(Selected|B)$ given two groups A and B. Another fairness metric is equalised odds which is achieved when the rate of selection is equal between groups for both qualified and not qualified individuals:

$$P(Selected|A \cap Qualified) = P(Selected|B \cap Qualified)$$

$$and$$

$$P(Selected|(A \cap !Qualified) = P(Selected|B \cap !Qualified)$$

Equalised outcomes is when the probability of being qualified is equal for both groups given an individual has been selected. The probability of being qualified should also be equal for both groups given an individual was not selected.

$$P(Qualified|A \cap Selected) = P(Qualified|B \cap Selected)$$

$$and$$

$$P(Qualified|A \cap !Selected) = P(Qualified|B \cap !Selected)$$

Statistical parity is used throughout this paper due to wanting to select a simple measure that focused on group fairness, not considering whether or not the person is qualified, thus the tells if the models are discriminating only based on sex and race.

Models

Two different models were chosen for the purpose of this paper, a logistic regression and a random forest. The logistic regression has high explainability and interpretability. While predictions of a random forest are possible to explain; the complexity and randomness in the construction of the trees and the features lowers the interpretability. For the logistic regression to work properly, the categorical variables are one-hot encoded, beside one value, which is left out to enable comparison. This value is therefore represented by zeros in all of the other values' columns. Whilst these features were one-hot encoded for the logistic regression, they were preserved in their original form for the random forest, since this type of model does not require columns to be in a one-hot encoded. In order to make the results more reliable, a five fold cross-validation was used to ensure the robustness for the models' results.

5 Analysis

Before the cross-validation could be performed the data were shuffled and split into 5 equal sized dataframes, where for each fold a dataframe served as a test set, the rest, training data. For each fold, the training set was used to fit a standard scalar and which was then used for transforming both the test and train set. Only the train set within each fold was used to fit the scalar to prevent information leakage. The training set was then re-projected to be orthogonal to the plane spanned by the protected features using the geometric re-projection methodology. This was regulated with the lambda value $\lambda \epsilon [0:1]$, which increased by 0.01 after each cross-validation was performed.

For each cross-validation the mean statistical parity and mean accuracy were calculated for each of the protected features. The coefficients for the logistic regression were also saved in the same manner, alongside the feature importance of the random forest model. It is important to note, that feature importance and coefficients were also averaged across the five folds.

The hyperparameters of the random forest model were set to 200 trees, with a max depth of 5 and a minimum sample split of 10. For the logistic regression the maximum number of iterations was set to 1000, such that the model could converge, furthermore the class weight was set to be balanced, thus weighting the classes proportional to their occurrences in the dataset.

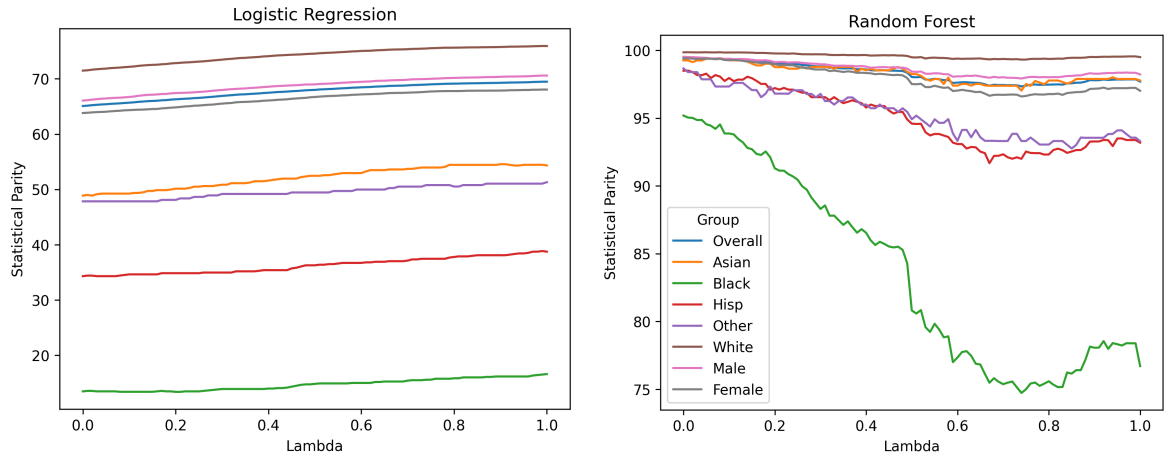
6 Results

This section will state the impact of the varying lambda values on the accuracy, statistical parity, and feature importance of the models.

It was found that the accuracies for the random forest model were higher than for the logistic regression (Appendix: B). Furthermore, the impact of the re-projection on the protected feature sex,

was higher for the logistic regression. Whilst the accuracies for the logistic regression fell with lower levels of lambda, they almost stayed unchanged for the random forest model. The same picture is painted when looking at the races, except for *Black* and *Other*, where the random forest model had a significantly lower accuracy for the re-projected data than for the original. Furthermore, another interesting finding was that the accuracy for the racial group *White* stayed almost unchanged at around 92%.

For the statistical parity, it can be seen that the re-projection had a equalising effect on the random forest model for all of the protected features (Figure: 3b). While the sex became more equal, and is almost the same, it can be seen that the predicted pass rate of the race *Black* increased significantly, from approximately 77% to 95% for the random forest model, making the gap in selection rates more narrow between *Black* and the other races. Thus, the re-projection had had a positive effect on the statistical parity for both sex and race. However, when looking at the logistic regression, the re-projection had almost no impact on any of the protected features, thus the gap between the different protected features was not narrowed (Figure: 3a).



(a) Logistic Regression

(b) Random Forest

Figure 3: Line plots for statistical parity, for all groups, for both logistic regression (a) and random forest (b). Note that the axis are not synchronized.

The feature coefficients for the logistic regression showed that the importance of the schools' mean ranking coefficient fell as the data became more unbiased, however, it was still the feature with the biggest impact on prediction for the $\lambda = 0$ model, followed by LSAT score (Appendix: C). For the random forest, the school mean ranking and then the LSAT score, were also the feature with the highest importance. However, both features' importance decreased as lambda decreased and as the data became less biased. Interestingly, the *full-time* study feature became more important as the data became less biased; furthermore, it was the only feature that significantly increased.

Overall, the re-projection had a negative impact on the accuracy, but no effect on the statistical parity for the logistic regression. Whereas, for the random forest model, re-projection had a more

positive impact in terms of the statistical parity but a slightly negative impact on the accuracies for all protected features except for the race *Black* where the impact was more negative than for the other projected feature values.

7 Discussion

This section will discuss the results, the impact of geometrically re-projecting the data, the explainability and interpretability of the models, and the potential impact such models can have on individuals and society.

7.1 Accuracy, Statistical Parity and Re-Projection

From the results, it could be seen that the geometric re-projection of the data had an impact on the accuracies for the protected features; especially for the logistic regression. This is expected, as the process involves removing some information about the protected features from the data. The geometric re-projection removes linear correlations between the features and the protected features, and since the logistic regression is a generalized linear model it therefore looses information about the protected features as it cannot capture non-linear correlations in the data. However, even though re-projection is independent of the target variable, it may influence other linear correlations between the features (e.g. the LSAT score) and the target variable. This argument could further be substantiated by the fact that the geometric re-projection did not influence the accuracy as much for the random forest, since the random forest can capture non-linear relationships between the features.

However the random forest's accuracy for the race *Black* was lowered as λ decreased, this could be due to many reasons, such as the ratio of *Black* people passing the bar being far lower than for other racial groups, at around 60%. This may be due to the large increase in the selection rate of *Black* individuals for the more re-projected models. Therefore the model makes more false positive errors for this group, likely due to the bias being removed from the dataset. However, the accuracy for this group was still higher for the random forest than for the logistic regression.

When looking at the statistical parity for the different groups, the geometric re-projection increased the selection rates of groups for the random forest. Furthermore, it narrowed the gap between the protected features; thus making the rate of predictions that one would pass the bar more equal across the protected features. While de-biasing the data lowered the selection rates for all groups, it did not influence the statistical parity between the groups for the logistic regression. This may be due to the low complexity of the model, limiting the information that it could extract from the features. Overall, geometric re-projection made the random forest model more fair, while almost preserving the same accuracy level for most of the groups, while the linear regression model struggled both to increase the fairness between the protected groups, and to preserve the same level of accuracy.

One of the drawbacks of geometric re-projection is, as mentioned above, that it only removes

linear correlations, leaving non-linear relationships between the features and the protected features in the data. This has an impact on generalized linear models, which cannot capture these non-linear relationships, therefore could potentially make it more difficult for these models to find a pattern in the data; like the logistic regression made in this paper. Another side-effect of removing these linear correlations, is that correlations between the target and the features also may be affected, thus having a negative impact on the accuracy for the models, since the data have lost some of its underlying information about the features and the target. Therefore the λ parameter plays an important role, as it gives control over this trade-off.

7.2 Explainability and Interpretability

The logistic regression models are a good choice when it comes to the explainability, since the coefficients for the different features can easily be gleaned from the model. These provide easy interpretability, showing which of the features impact the predictions most, providing both the direction and magnitude of each feature's impact. On the other hand, the logistic regression has some shortcomings in terms of capturing the complex non-linear relationships, here a random forest is a better choice since it can capture these patterns. However, a random forest model is less interpretable, since it by definition has some randomness in terms of how each of the trees are constructed and which features are used; without being able to explain why this combination of features and trees works. On the other hand, it is possible to dive into each of the trees, thus making it possible to interpret the workings of the model. However when the number of trees are high, this investigation can take some time and be confusing, where it for the logistic regression would be easier to find the values of the coefficients and to interpret the working of the model.

It is possible to extract the overall feature importance of the random forest, using the Mean Decrease in Impurity (MDI), thus making it possible to explain which features has the most impact on the prediction given by the random forest. However, this calculation has been showed to be biased toward features with many different values [4]. That being said, the MDI feature importances for the random forest model aligned well with the coefficients of the logistic regression, thus for the purpose of this study this bias was found to have a minimal impact on the analysis and results.

Overall, there is a trade-off between the explainability and interpretability and the accuracy of the model. Where one should consider if the model should be explainable, such that it is possible to justify whether or not a university accepts an applicant, or whether one weights the accuracy higher. Investing in explainability often requires more resources; for example time and money. This trade-off will be discussed further in the following sections.

7.3 Limitations and Bias

The LSAC dataset was collected using a self-report questionnaire from the autumn semester starting class of 1991. Data from 68% of the 40,000 students was included. Self-reports can introduce bias in multiple ways, one of these is social desirability bias, where individuals change their answers to reflect a more social acceptable version of themselves. However, this effect was limited by the the use of a comprehensive consent form detailing stringent data privacy measures that were in place. Sample bias may also play a role, participation in the survey was voluntary, but certain groups such as students of colour or low-performing students may have felt less comfortable than others in participating in the study. Thus the data may not have been fully representative of the first semester law students from 1991. Furthermore, the data was collected over thirty years ago, at the point that this paper is being written. Temporal generalisability may therefore be low, as society has changed.

7.4 Future work

Further analysis could have been carried out by comparing fair-PCA of the data alongside the geometric re-projections[5]. Furthermore, explainability of the models could have been explored at a more local level using Shapley Additive explanations (SHAP) values, to give more context to the existing global explanations of the feature importances of the models. SHAP values provide context to local feautre imporance, and can thus provide a fine-grained and instance specific explainability for a models classification.

7.5 Reflections

Good old-fashioned AI vs New AI

Good old-fashioned AI (GOFAI) is where symbolic representations of an input are manipulated by the AI, following a specific set of instructions to generate an output. New-AI is less explainable, it works more as a correlation detection mechanism, where deep learning architectures enable the models to see patterns that humans can not necessarily understand or interpret. It is a challenging task to remove bias from new-AI models, as they have the ability to draw complicated and non-linear correlations from real-world data. The features of the LSAC dataset for example, have variables that linearly correlate with some of the protected features; however, there exist non-linear correlations, which models, such as the random forest, are able to detect. This means that bias may remain in the models.

Plato's Meritocracy and Liberal Democracy

Plato's idea of Meritocracy centers around the idea that individuals should be selected to hold positions of power based on their abilities and virtues as decided by experts. Each year professors, professional admission staff and AI algorithms, judge and select whom to admit, in a process reflecting Plato's vision of experts choosing the new generation of experts. Following this philosophy, in

the case of law school admissions, one should just choose those individuals that are most able and therefore most likely to be successful and pass the bar. Moreover, putting the needs of the society higher than the needs of the individual, by removing freedom of choice. In contrast, a keystone of liberal democracy is that one should not discriminate based on certain protected traits of a person. According to these ideas, the role of the government should be to provide equal opportunities for individuals to choose their path in life themselves. An individuals background should not determine their future; in practice fair admissions may therefore involve implementing measures, so that more equal statistical parity is achieved between groups.

The discussion of Plato's meritocracy compared to liberal democracy, raises further questions about fairness and justice. Fairness is a complicated and multifaceted concept, however in general it refers to the equal opportunities, treatment, and outcomes of different individuals. Justice is in contrast related to application of the rules and laws of a society. With the rise of algorithms used to screen applications for life changing opportunities such as law school, it is more important than ever to consider how this impacts fairness. Removing protected features is often not enough, since other information can be indirectly linked to these protected features. However, due to correlations with features such as race these AI algorithms may make bias decisions. Even if de-correlation methods are utilised, New-AI's often non-linear nature means that linear de-correlation is sometimes not very effective.

8 Conclusion

In conclusion, this paper found that the geometric re-projection of the LSAC dataset, and therefore linear-de-correlation, influenced not only the accuracy of the models, but also the statistical parity for the protected features of race and sex. Furthermore, it was found that the different lambda values influenced sex and race in different ways, for both models. The geometric re-projection lowered the accuracies of both models. However the random forest model, was found to be more robust, thus the accuracy did not drop as much as for the logistic regression. Moreover, the statistical parity was found to increase with lower levels of λ, narrowing the gap between the groups. The ethical implications of deploying these types of models still warrants continuous future investigation, since the fairness of an application acceptance is still subjective for each faculty of the universities. Furthermore, decisions regarding fairness often involve a trade off between the wishes of an individual and society; should a applicant be accepted based solely on their qualifications, or should one enforce diversity by implementing policies such as gender quotas.

References

[1] Linda F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998. URL https://api.semanticscholar.org/CorpusID:151073942.

[2] Cole Claybourn. Is ai affecting college admissions? *U.S News World Report*. URL https://www.usnews.com/education/best-colleges/articles/is-ai-affecting-college-admissions.

[3] Yuzi He, Keith Burghardt, and Kristina Lerman. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, page 279–285, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375864. URL https://doi.org/10.1145/3375627.3375864.

[4] Permutation importance vs random forest feature importance (mdi)¶. https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html. Accessed: 2024-05-20.

[5] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair pca for fair representation learning. 2023.

Appendices

A Description of raw dataset

Feature Name	Description	Kept	Reasoning
ID			Not relevant
bar1	Passed bar 1st Attempt	yes	Chosen Target Variable
bar1_yr	Passed bar 1st Attempt (detailed)		not binary
bar2	Passed bar by 2nd Attempt		High class imbalance
bar2_yr	Passed bar by 2nd Attempt (detailed)		not binary
pass_bar	Passed bar by 2nd attempt		Duplicate of bar2
bar	When passed bar if graduated		not binary
bar_passed	Like pass_bar but vaires, unknown reason		Unclear data
dnn_bar_pass_prediction	A prediction of the bar exam result		Not relevant
sex	1 or 2		Duplicate of male
gender	female/male		Duplicate of male
male	0 or 1	yes	Protected attribute
race	7 unspecified groups		Unclear data
race1	black, white, asian, hispanic, other	yes	Protected attribute
race2	black, white, other		Duplicated in race1
hisp	One hot encoded		Duplicated in race1
asian	One hot encoded		Duplicated in race1
black	One hot encoded		Duplicated in race1
other	One hot encoded		Duplicated in race1
ugpa	GPA, Undergraduate		Duplicate of gpa
zfygpa	GPA, First Year of Law School		Not Available during Admission
zgpa	GPA, Cumulative Law School		Not Available during Admission
gpa	Undergraduate GPA	yes	
lsat	LSAT Score	yes	
decile1b	No description of data		Unclear data
decile3	Law school rating in third year	yes*	averaged with decile1
decile1	Law school rating in first year	yes*	averaged with decile3
fulltime	Whether a student was enrolled fulltime	yes	
parttime	Whether a student was enrolled parttime		Duplicated in fulltime
fam_inc	Family Income Quintile (FIQ)	yes	
age	Negative values?		Unclear data
DOB_yr	Year of birth	yes	Used to calculate age
grad	If an individual graduated law school		Not Available during Admission
dropout	If an individual dropped out of law school		Not Available during Admission
tier	Tier of law school by quintile		Correlation: decile1, decile3
index6040	unsure		Unclear data
indxgrp	unsure		Unclear data
indxgrp2	unsure		Unclear data
cluster	unsure		Unclear data

B Accuracy of the models

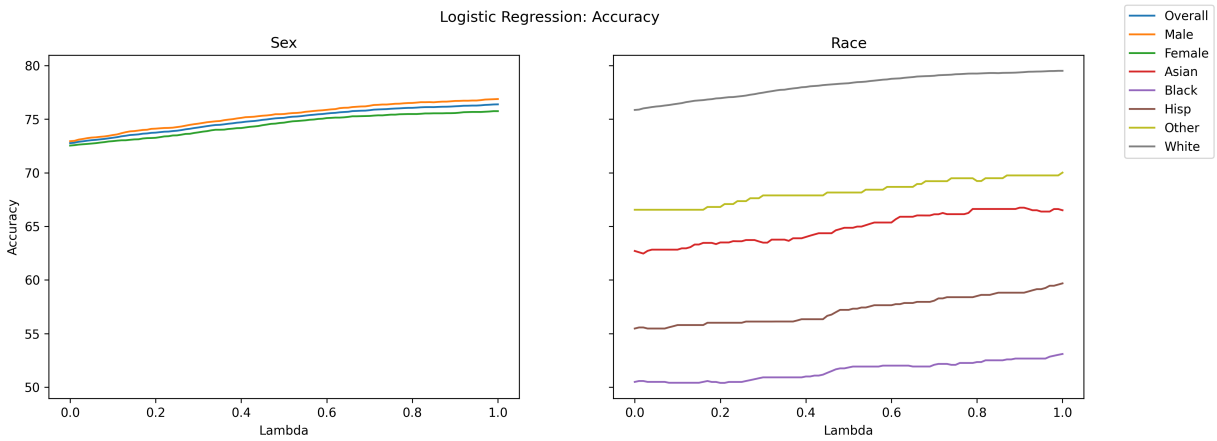


Figure 4: Accuracy for the logistic regression

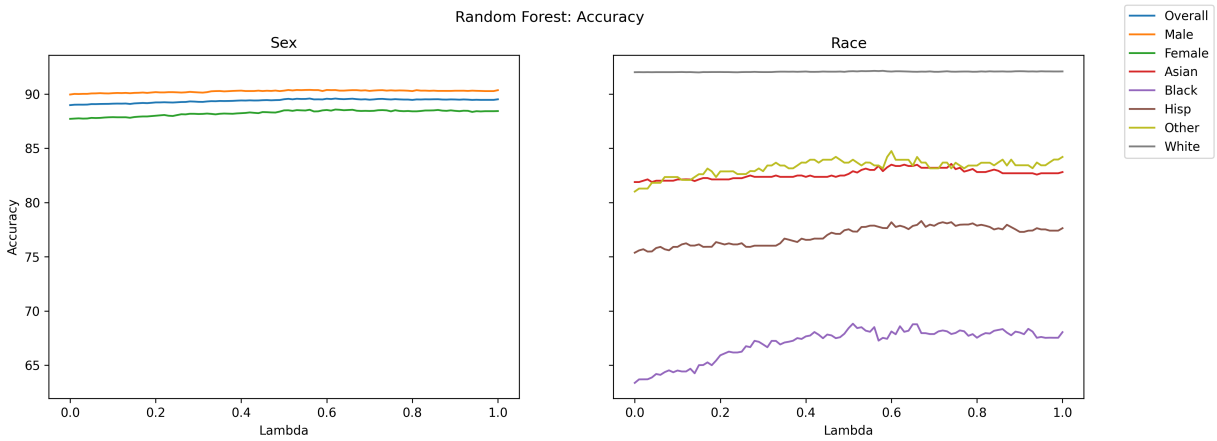


Figure 5: Accuracy for the random forest

C Feature importance of the models

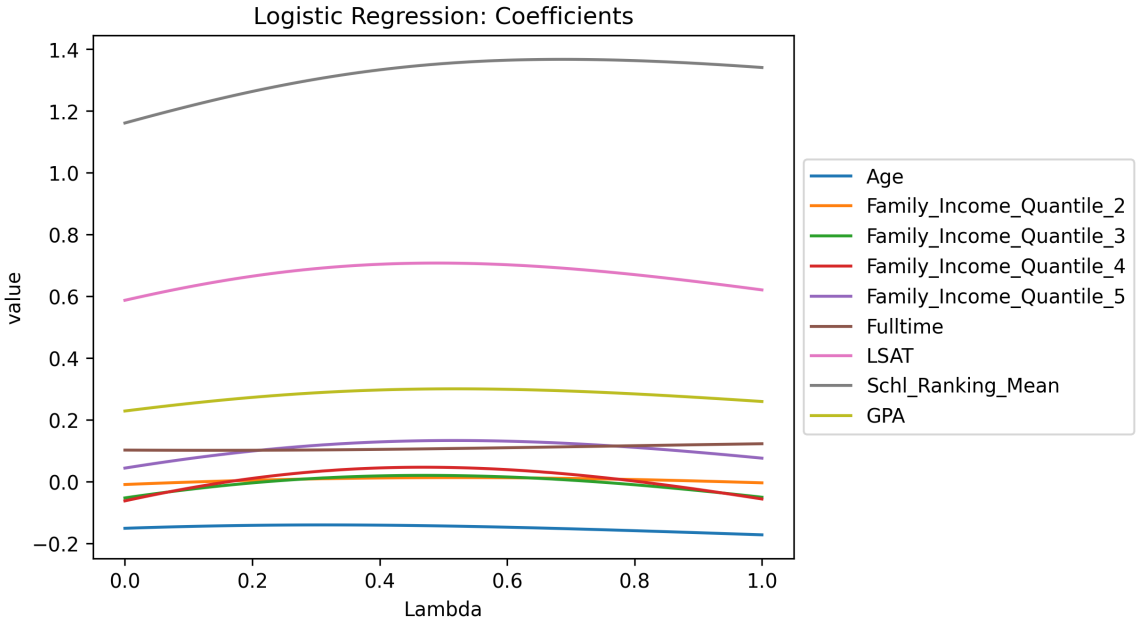


Figure 6: Feature coefficients for the logistic regression

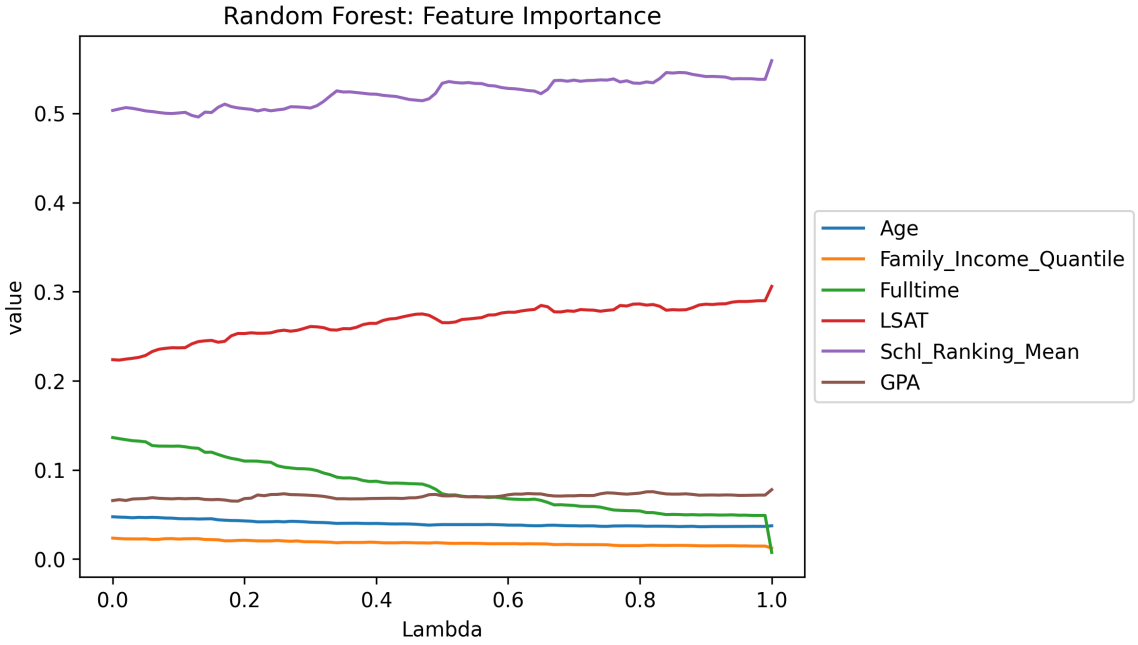


Figure 7: Feature importance for the random forest