**Who Pass the Bar?**

**2022 Fall Introduction to Data Science**

**Team 8: Brunda Mariswamy, HaeLee Kim, Mahikshit Kurapati**

**Chapter 1: Introduction**

A Law School Admissions dataset from the Law School Admissions Council (LSAC). From 1991 through 1997, LSAC tracked some twenty-seven thousand law students through law school, graduation, and sittings for bar exams. The result was the most comprehensive database that exists on the demography, experiences, and outcomes of a large cohort of aspiring lawyers. While the data has important limitations, it is a unique and very valuable source for studying a range of issues related to legal education. The dataset was originally collected for a study called 'LSAC National Longitudinal Bar Passage Study' by Linda Wightman in 1998.Data set contains 22407 observations with 39 variables. We collected the data set from Kaggle.

**1.1 SMART Questions:**

Idea: Predict whether a student will pass the bar, based on their Law School Admission Test (LSAT) score and undergraduate GPA.

**SMART Questions:**

1. **What contributions affect the LSAT score significantly?** (Models: LR, Decision Regression Tree, PCR)
2. **What contributions significantly affect clearing the bar exam? (**Models: Logit Regression, KNN)

**1.2 Preparing for EDA**

* Dropped NAs from the data set.
* Dropped columns with duplicate column information.

Text

Description automatically generated

* Converted data types of variables

A screenshot of a computer

Description automatically generated with medium confidence

**1.3 Summary of the data set**

* **Dependent Variables:** LSAT score, Bar Pass (0: Failed,1: Passed)
* **Independent Variables:** Undergraduate GPA, Law School GPA, Full time, Family income, Gender, Race, Law School Tier.

**Text

Description automatically generated**

**A black screen with white text

Description automatically generated with low confidence**

**1.4 Models and Graphs used in the Analysis**

For EDA we have used ANOVA and for answering the SMART question we used Linear Regression, Decision Tree Regression, PCA, Logit Regression, KNN models. We have used Histogram, Boxplot, Correlation Matrix in the whole analysis.

**Chapter 2: EDA**

**2.1 General Analysis of Data set**

1. **Total percentage of students who passed the bar.**

Chart, pie chart

Description automatically generated

From the pie chart we observe that 94.9% of the students have successfully cleared the bar exam.

1. **Total percentage of male/female candidates who appeared for bar exam.**

Chart, pie chart

Description automatically generated

Male candidates are highest in number of 56% while female candidates’ only 44% who appeared for bar exam.

1. **Total percentage of students who enrolled for bar exam who belong to different race.**

Chart, pie chart

Description automatically generated

From the data we observe that race1 category students are highest with 84% while race 5 being lowest number who appeared for exam.

1. **Total percentage of students who enrolled for full time course.**

Chart, pie chart

Description automatically generated

We see that 92.6% of students enrolled for full time course while only 7.4% students opted for part time.

1. **Total percentage of students who enrolled in different tier university.**

Chart, pie chart

Description automatically generated

From pie chart we observe that most of the students enrolled to tier 3 and tier 4 university while only few admitted to tier 1 which is having low ranking.

**2.2 Normality Check (Histogram) for LSAT, GPA, Undergraduate GPA**

From the Histograms plotted for LSAT, GPA, UGPA we observe that all 3 histograms are slightly left skewed.

Chart, histogram

Description automatically generated

The distribution of LSAT score as average LSAT Score of 36, with 48 being highest score while 11 is the lowest score.

Chart, histogram

Description automatically generated

The Distribution of GPA as average GAP of 3.2 with 3.9 student’s highest GPA and 1.5 lowest GPA.

Chart, histogram

Description automatically generated

The distribution of Undergrad GPA as average GPA of 3.2 with student highest UGPA of 3.9 while 1.5 minimum GPA.

**2.3 Multiple Boxplot & ANOVA Test**

Box plot and ANOVA test to check the significance between the categorical variables.

**LSAT**

1. **LSAT Score by University Tier**

Chart, box and whisker chart

Description automatically generated

A black screen with white text

Description automatically generated with low confidence

For the distribution of LSAT score and different university tier we see that the tier 1 university to tier 6 has different mean and the value of p is less than standard alpha value 0.05, Hence we say there is significance difference between LSAT score and university tier.

1. **LSAT Score by Race**

The distribution of LSAT score and different race group we see that mean of LSAT score is different for race group 1 to race 5 and from ANOVA test results we observe that p value is less than alpha value, hence we can see significance difference between the race groups.

Chart, box and whisker chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

1. **LSAT Score by Gender**

Chart, box and whisker chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

The distribution of LSAT score and gender we observe that the mean of LSAT score for both male and female is almost same, even though the value of p is very less but we see no significance difference between the gender variable.

**GPA**

1. **GPA by University Tier**

Chart, box and whisker chart

Description automatically generated

A black background with white text

Description automatically generated with medium confidence

For the distribution of GPA and different university tier we see that the tier 1 university to tier 6 has different mean and from ANOVA test results we see that the value of p is less than standard alpha value 0.05, Hence we say there is significance difference between GPA and university tier.

1. **GPA by Race**

Chart, box and whisker chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

The distribution of GPA and different race group we see that mean of GPA is different for race group 1 to race 5 and from ANOVA test results we observe that p value is less than alpha value, hence we can see significant difference between the race groups.

1. **GPA by Gender**

Chart, box and whisker chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

The distribution of GPA and gender we observe that the mean of GPA for both male and female is almost same, even though the value of p is very less but we see no significance difference between the gender variable.

**Undergraduate GPA**

1. **Undergraduate GPA by University Tier**

Chart, box and whisker chart

Description automatically generated

A black background with white text

Description automatically generated with medium confidence

For the distribution of UGPA and different university tier we see that the tier 1 university to tier 6 has different mean and from ANOVA test results we see that the value of p is less than standard alpha value 0.05, Hence we say there is significance difference between UGPA and university tier.

1. **Undergraduate GPA by Race**

Chart, box and whisker chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

The distribution of UGPA and different race group we see that mean of UGPA is different for race group 1 to race 5 and from ANOVA test results we observe that p value is less than alpha value, hence we can see significance difference between the race groups.

1. **Undergraduate GPA by Gender**

Chart, box and whisker chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

The distribution of UGPA and gender we observe that the mean of UGPA for both male and female is almost same, even though the value of p is very less but we see no significance difference between the gender variable.

**2.4 Correlation Table**

We have used three types of correlation plot considering all variables from the data frame to check the correlation between the variables.

From correlation plot we observe that race is highly correlated with LSAT score and pass bar variable.

1. Correlation matrix

A picture containing background pattern

Description automatically generated

1. Scatter Matrix

A picture containing text, shoji

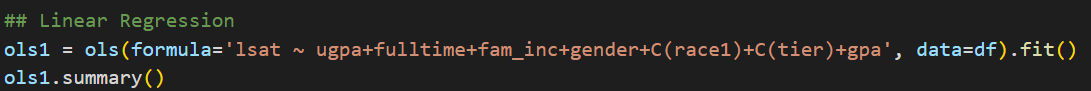
Description automatically generated

**Chapter 3: Prediction Analysis**

**3.1 Statistical OLS Regression**

We used the statistical OLS model to check what factors affect the LSAT score.

* **Dependent Variables:** LSAT score
* **Independent Variables:** Undergraduate GPA, Law School GPA, Full time, Family income, Gender, Race, Law School Tier.



The result shows that dependent variables statistically affect the LSAT score. In other words, the higher undergraduate GPA, the higher law school GPA, the higher school tier, and the higher family Income are positively related to the higher LSAT score. Also, male students and full-time students tend to get higher LSAT scores. Based on the adjusted R-squared, 35.7% variability in LSAT scores can be explained by dependent variables in this linear model.

Text

Description automatically generated

**3.2 Sklearn (LinearRegression)**

To find out how the previous model matches the current dataset, we split the dataset into a train set (80%, 17588) and a test set (20%, 4397) and used sklearn (LinearRegression). After the analysis, the Sklearn result shows the training score as 0.277 and the test score as 0.287.

Text

Description automatically generated

Lastly, we checked the cross-validation for Linear Regression Model (CV = 5) and the result mean was 0.273.

Text

Description automatically generated

**3.3 Decision Tree Regression**

Decision Tree Regression trains a model in the structure of a tree to predict the future data to produce the continuous output, a set of numbers or values. The outcome predicts the average value of the target variable, in our model the target variable is the LSAT score. Decision Tree Regression is split by variance, standard deviation, etc. instead of entropy and is commonly evaluated by RMSE (Root Mean Squared Error) or MAE (Mean Absolute Error).

The results show that the test set RMSE of the regression tree is 4.84, the RMSE of the linear regression test set is 4.65, and the RMSE of the regression tree test set is 4.84

.**Text

Description automatically generated**

The branches are created by cut points dividing the space into regions and those cut points are determined by minimizing the MSE of both new regions. The tree below shows that the school tier lower than 2.5 has an average LSAT score value of 31.933 and the school tier over 4.5 and the undergraduate school GPA over 3.15 will have an average LSAT score of 40.418.

**Diagram

Description automatically generated**

**3.4 Principal Component Analysis (PCA)**

PCA is an unsupervised learning method, a non-parametric statistical technique to perform dimensionality reduction. The reason for focusing on reducing the dimension is that the high dimensionality will bring model overfitting and reduce the generalization ability except for the training set.

**Chart, line chart

Description automatically generated**

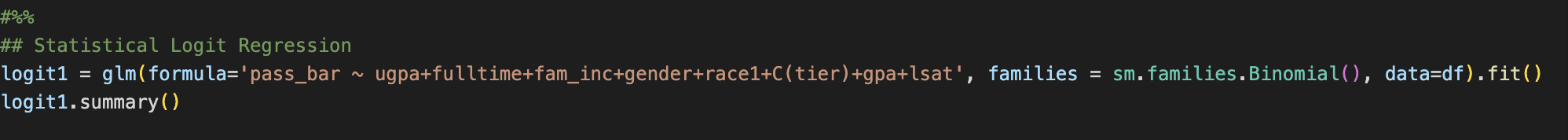
First, we scaled the data to fit the PCR model. Then, from the visual representation of the PCR values, we found out that we can reduce the computation and achieve the same results by only using the first six variables. The result suggests that the first 6 variables are most optimal for training the regression models and wouldn't make much of a difference even if we didn't include the 7th variable. The RMSE was reduced to 4.648 (using Linear Regression).

**Chapter 4: Classification**

**4.1 Statistical Logit Regression**

We used the statistical Logistic Regression model to build an algorithm to classify whether a person passed the bar.

* **Dependent variables:** Pass Bar (0: Fail, 1: Pass)
* **Independent Variables:** Undergraduate GPA, Law School GPA, Full time, Family income, Gender, Race, Law School Tier.



Text, calendar

Description automatically generated

In our logit regression, we can interpret that the logit-value > 0 means more likely pass the pass than fail. The result shows that independent variables statistically affect the classification of passing the bar. In other words, the higher undergraduate GPA, higher law school GPA, the higher school tier, and the higher family Income are positively related to the higher chance to pass the bar exam.

**4.2 Sklearn(Logistic Regression): Before SMOTE**

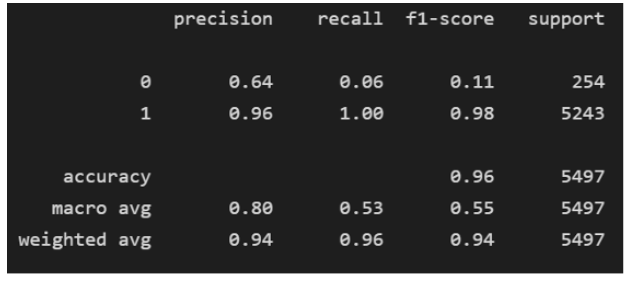
For Logistic Regression in sklearn, we split the data into 80% training and 20% test sets.

Then we trained the Logistic Regression using the following code.

**Text, letter

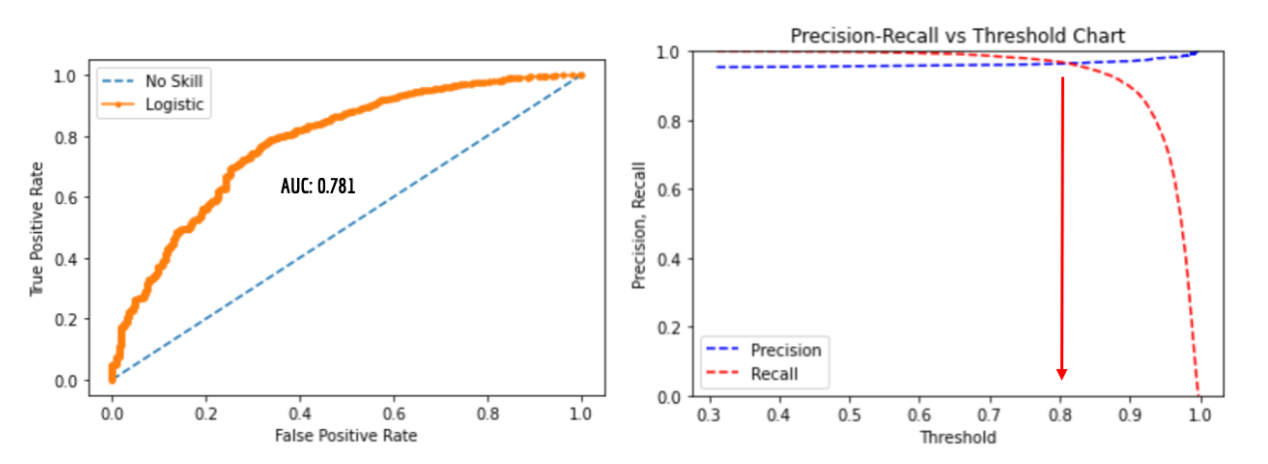
Description automatically generated**

We got accuracy 96% before SMOTE

****

In our logit regression, we can interpret that the logit-value > 0 means more likely pass the pass than fail. The result shows that independent variables statistically affect the classification of passing the bar. In other words, the higher undergraduate GPA, higher law school GPA, the higher school tier, and the higher family Income are positively related to the higher chance to pass the bar exam.

Then, we checked the AUC score and got the following result:

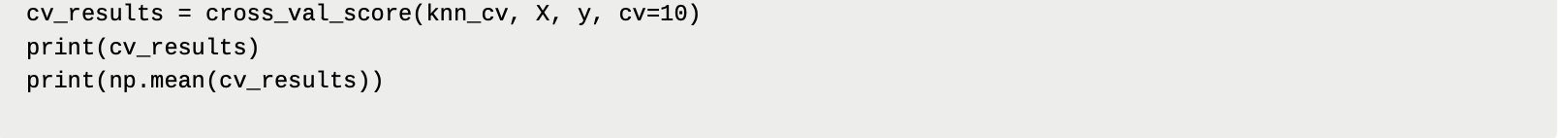


AUC is an aggregated metric evaluating how well a logistic regression model classifies positive and negative outcomes at all possible cutoffs. AUC also means for evaluating predictive performance of a model. The result shows the AUC score as 0.781 ~ 78.1%, which means that the model can differentiate between different classes adequately.

The higher threshold, the higher the precision, but the lower the recall: the deal threshold setting is the highest possible recall and precision rate. From the second graph, we can see that the threshold is 0.8. The higher the threshold, the higher the precision, but the lower the recall: the deal threshold setting is the highest possible recall and precision rate.

**4.3 KNN with Cross-Validation: Before SMOTE**

For the K-Nearest Neighbors, we used the k value as 7 and after 10 folds of cross-validation, we got an R2 score of 94.6%. We also split the dataset into Train (80%) and Test (20%) to use sklearn (KNeighborsClassifier).





Graphical user interface

Description automatically generated with low confidence

Here, we can see the error rate decreases as the k-values increase.

**A picture containing graphical user interface

Description automatically generated**

**4.4 SMOTE**

We use SMOTE here because the dataset is highly imbalanced. While there are 97% of students passed the bar, there are only 3% didn’t. This makes the classification biased, and while it predicts well the training and test data, the model won’t be able to handle data from external sources.

Then, we implemented SMOTE using the following code.

**Text

Description automatically generated**

**4.5 Sklearn(Logistic Regression): After SMOTE**

While we used the same parameters, Logistic Regression after running SMOTE on the data performs unsatisfactorily. This is probably because the pseudo data created by SMOTE is making the model predict wrong in some cases.

**Calendar

Description automatically generated**

**4.6 KNN: After SMOTE**

We can see in this error plot that after SMOTE, the error increases as the k value increases. As we are balancing the imbalanced dataset, the number of pseudo data points are increased. Hence, the accuracy of the model decreases however the value of k is 3.

**Chart, scatter chart

Description automatically generated**

Although k=1 shows the lowest error rate, it can overfit the model. For this reason, we got k=3 and the accuracy as 86% after SMOTE.

**Text

Description automatically generated**

As we know, SMOTE amplifies our data by creating pseudo data to balance an imbalanced dataset. By doing this, the score of our model has been drastically reduced as it tries to predict outcomes using pseudo data that is not practically real.

**Chapter 5: Conclusion**

**5.1 Analyses for Question 1. Prediction Summary Table**

Table

Description automatically generated

**5.2 Analyses for Question 2. Classification Summary Table**

**Graphical user interface, application, Teams

Description automatically generated with medium confidence**

**5.3 Final Summary**

Based on our finding for Q1, Undergraduate GPA, School Tier, Law school GPA, and Family Income positively relate to LSAT scores.

* Higher tier schools, Higher LSAT score
* Higher Undergraduate GPA, Higher LSAT score

Also, the findings for Q2 shows that Higher Law school GPAs, Undergraduate GPAs, LSAT scores, and School tiers increase the chance of passing the bar exam.

* Higher Law school GPA, the higher chance of passing the bar exam

Throughout the analyses, demographic differences were found in different gender and race groups. For example, full time, male, higher family income, and specific ethnic group show better performance in LSAT & Bar pass.