PROJECT 2 DSAN 780

Predictive Analytics, Spring 2024

Ferris State University, Instructor: Dr. Elies Kouider

.

Zhen Li, George Butvilas, John McCrackin

April 24, 2024

Abstract:

This work is related to a dataset provided by a customer concerning its analysis using data science and analytic methods to predict University admission. The dataset is comprised of more than 17,300 cases (students) and 12 fields. The fields to be utilized for the study are (1) the Applicant number, (2) the first parent's education level, (3) the second parent's education level, (4) student gender, (5) whether or not the student was White, (6) whether or not the student was Asian, (7) the student's high school GPA, (8) SAT or ACT score, (9) what colleges at the University the student was applying to, the College of Business and Economics, the College for Math and Science, or the College of Arts and Letters, (10) whether or not the student was admitted, (11) whether or not they were enrolled, and (12) and the college GPA. For the analysis, Applicant number, Enrolled, and College GPA are dropped. The customer is interested in the development of a statistical modeling tool that can predict whether the student will be admitted to the University. Our report follows the "CRSIP-DM" format (Cross Industry Standard Process for Data Mining) that follows the steps (1) Business understanding, that is what is the business need the customer is asking to be addressed, (2) Data understanding – an analysis of the data the customer has provided including its description, exploration, and an analysis of its quality, (3) Data preparation – selecting fields to be modeled, formatting and cleaning the data, and derivation of new fields as may be required to ensure model robustness, (4) Modeling – the modeling approach(es) such as regression models - linear or non-linear, neural nets, or other algorithms, and model building based on selecting a subset of models to evaluate, (5) Evaluation – investigating the strength of the models based on their predictive capability under various assessment criteria as well as their comparison to each other's, and (6) Deployment – final model recommendation and support for its use as the best model to serve the purpose. We evaluated multiple models, investigating the relationship between the inputs and the target variable using both linear and logistic regression models. The model of focus, however, was a Logistic Regression model based on the output – binary, admission to the university or not. As well we evaluated a plurality of input scenarios in the models and vetted them through statistical measurements. Once our best Logistic Model was developed, we compared it to a Neural Network, KNN, C&RT, and Random Forest model. In the end, we were able to validate what we think is a robust Logistic Regression model to predict University admission. The software used in this analysis is IBM SPSS 14.0 modeler and Microsoft Excel.

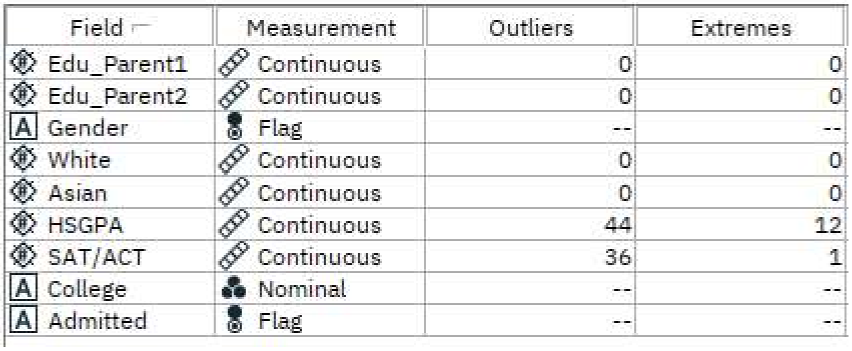
Part 1 - Business Understanding:

We have been provided with the dataset “College Admissions.” Our aim is to utilize predictive analytic techniques to determine what inputs from the dataset will best enable the forecast of whether a new student is admitted to the University. The dataset contains 17,399 cases (potential students). There are 12 fields in the dataset for us to utilize for our predictive model. The target field from the dataset is “Admissions.” The inputs we will use are: “Edu_Parent1,” “Edu_Parent2,” these inputs contain education levels for the students’ parents, student gender, student ethnicity, i.e., “White” or “Asian,” student GPA and SAT, the specific college within the University, i.e., Arts & Letters; Business & Economics or Math & Science, “Enrolled,” and “College-GPA.” We will be dropping fields from the study if they are deemed unnecessary. The main modeling technique we will use is Logistic Regression. The output of such a model is binary – that is, the prediction will be a “yes” or “no” to the admission question. As a course of our analysis, we will evaluate other models in addition to a logistic model and make some comparisons between them. The measurement of our success will be a robust model developed using sound data science and analytics methods validated by statistical metrics such as, but not limited to correlation analysis, input and model significance, Nagelkerke R square, Chi-Square, Hosmer and Lemeshow tests, Durban-Watson scoring, and tests that ensure data is not underfit or overfit.

Part 2 - Data Understanding:

The data set we’ve been provided contains 17,339 cases (applicants). In all, there are 12 fields in the dataset. We found no missing values in any field. We did find one small error we attribute to a typographical mistake we remedied in case number 17,310 regarding the case’s college.

A summary of the data as initially received and audited for quality is provided in **Figure 1**. Note three fields have been removed: “Applicant,” “Enrolled,” and “College_GPA.” From the initial audit, we find outliers and extremes for “HSGPA” and “SAT/ACT.” We will discuss mitigating the effects of these in the preparation step.



Field	Measurement	Outliers	Extremes
Edu_Parent1	Continuous	0	0
Edu_Parent2	Continuous	0	0
Gender	Flag	--	--
White	Continuous	0	0
Asian	Continuous	0	0
HSGPA	Continuous	44	12
SAT/ACT	Continuous	36	1
College	Nominal	--	--
Admitted	Flag	--	--

Figure 1: Audit of initial dataset "College Admissions."

With detail provided by the customer, the inputs from the dataset are summarized, supported by visual analytics, in the associated figures. Descriptions and statics for the inputs are as follows:

- Admitted (**Figure 2**):
  - Yes – 5,323 students admitted.
  - No – 12,016 not admitted.
- Gender (**Figure 3**):
  - Female – 10,038.
  - Male – 7,301.
- White (**Figure 4**) – 9,165 applicants identify as White.
- Asian (**Figure 5**) – 3,004 applicants identify as Asian.
- College (**Figure 6**):
  - Arts & Letters – 6,964.
  - Business & Economics - 4,103.
  - Math & Science - 6,272.
- Inputs "Edu_Parent1" and "Edu_Parent2," education level. More than half of parents had a four year or a post graduate degree (**Figure 7**):
  - 1-no high school
  - 2-some high school
  - 3-high school graduate
  - 4-some college
  - 5-associate degree graduate
  - 6-four-year college degree graduate
  - 7-postgraduate
- HSGPA (High School GPA) (**Figure 8**). Mean GPA is over 3.5.
- SAT/ACT (**Figure 9**). Mean SAT/ACT is 1,165.

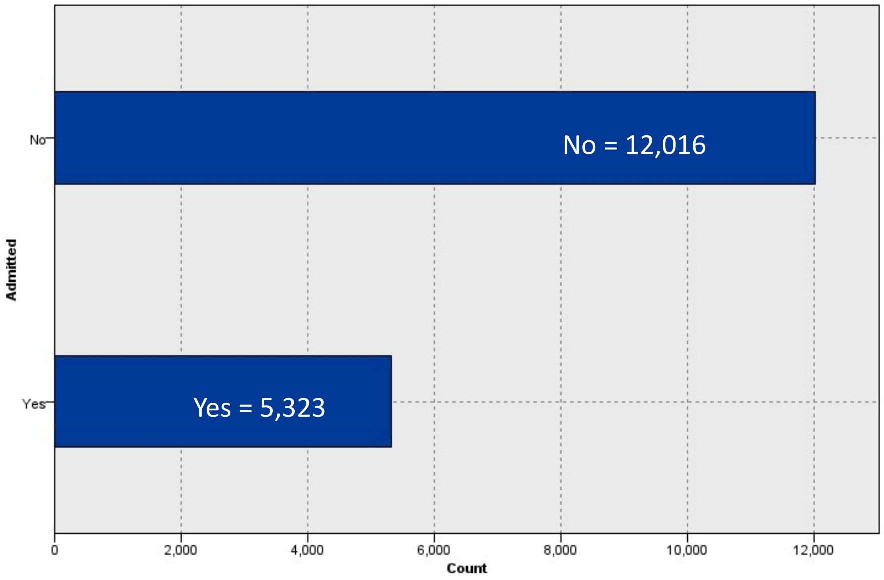


Figure 2: Admissions rates.

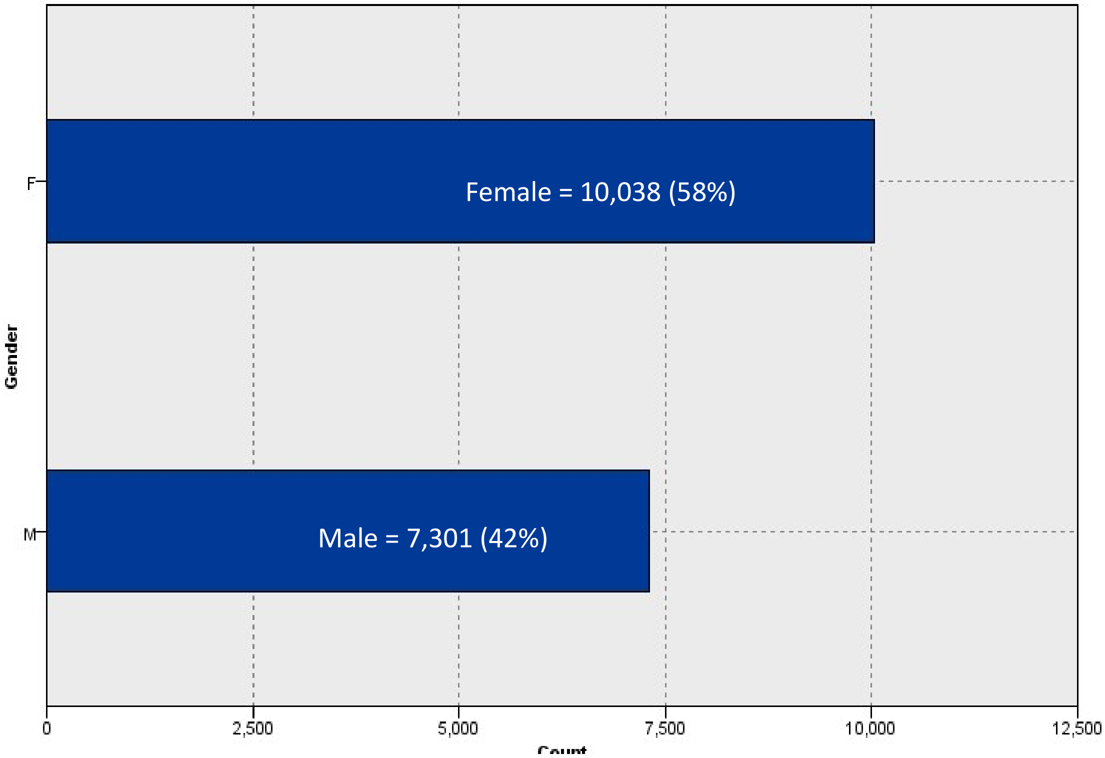


Figure 3: Gender statistics.

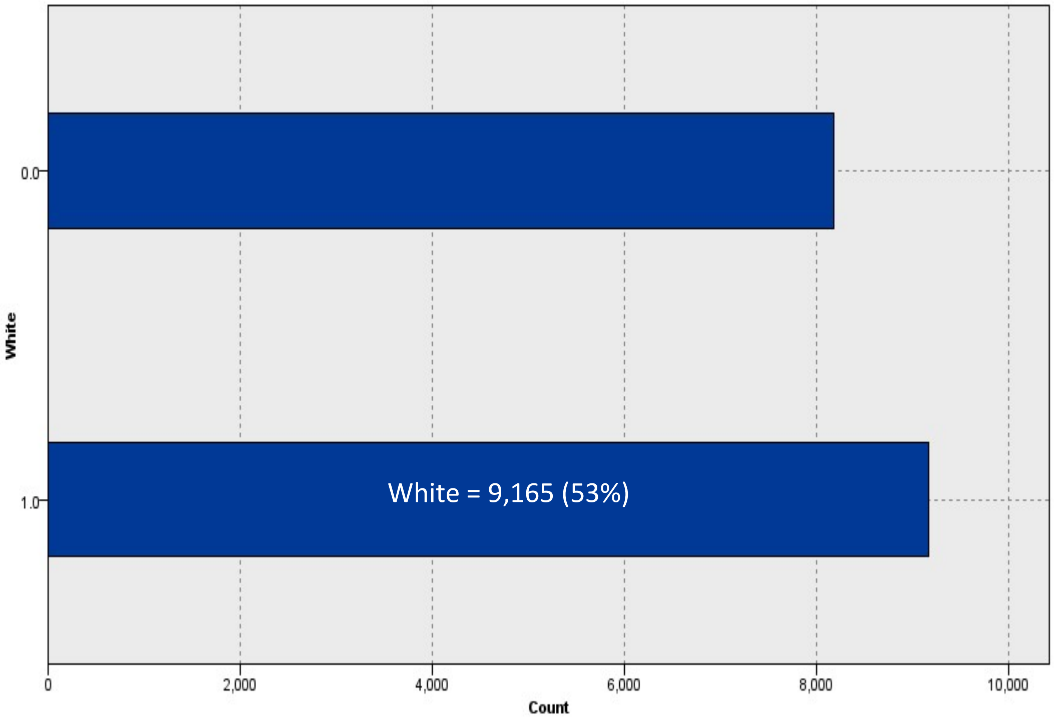


Figure 4: White applicants.

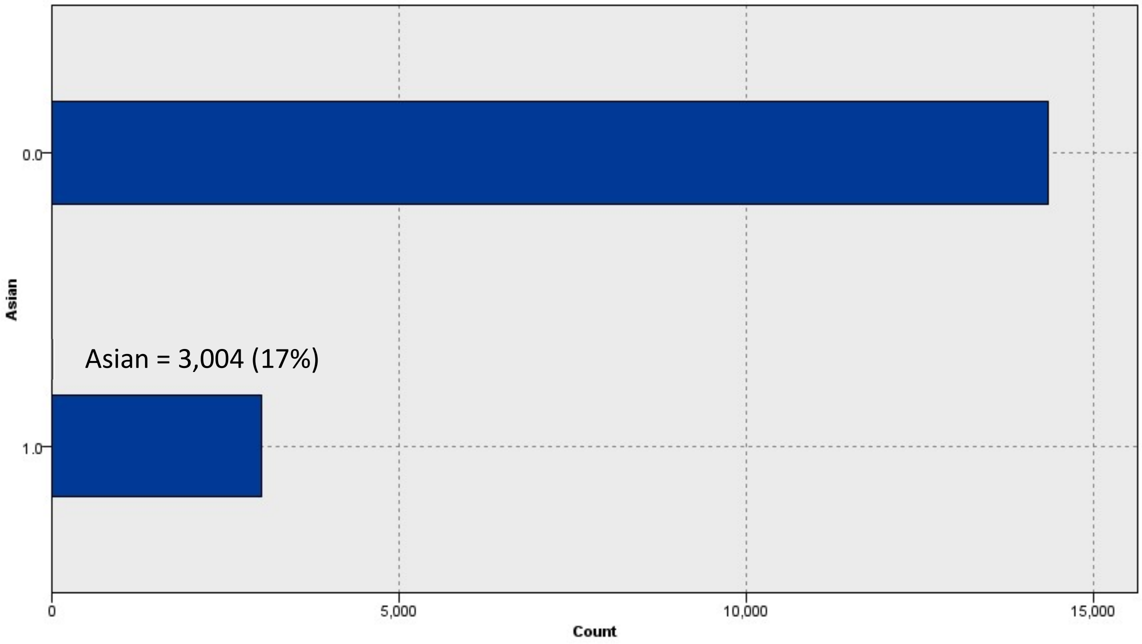


Figure 5: Asian applicants.

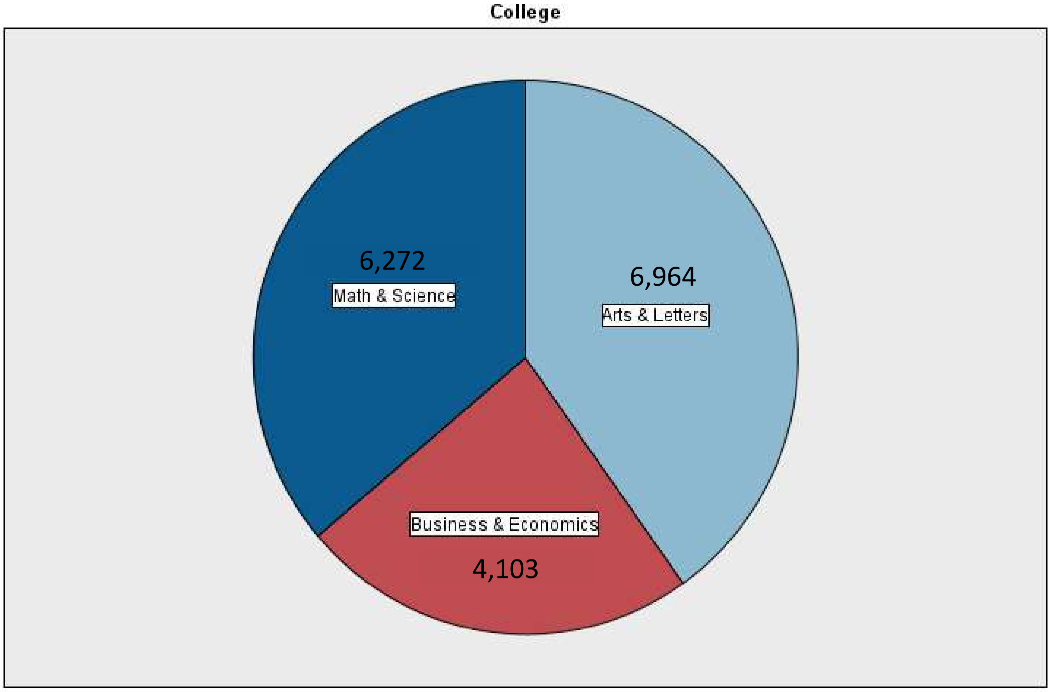
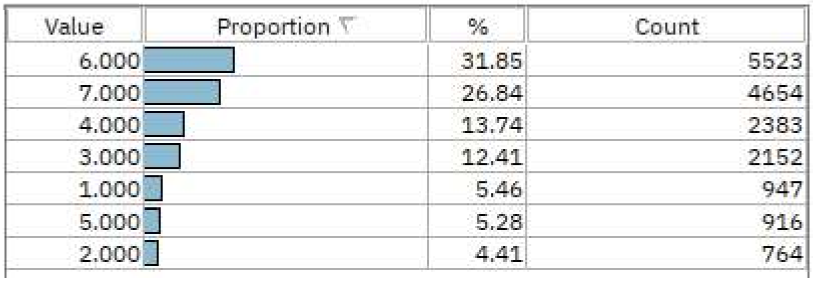


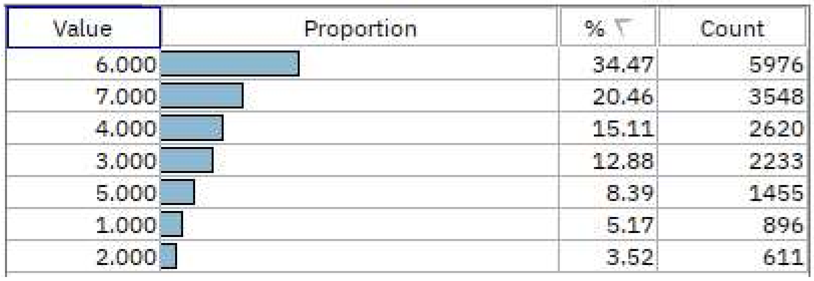
Figure 6: Colleges within the University.

Parent 1



Value	Proportion	%	Count
6.000		31.85	5523
7.000		26.84	4654
4.000		13.74	2383
3.000		12.41	2152
1.000		5.46	947
5.000		5.28	916
2.000		4.41	764

Parent 2



Value	Proportion	%	Count
6.000		34.47	5976
7.000		20.46	3548
4.000		15.11	2620
3.000		12.88	2233
5.000		8.39	1455
1.000		5.17	896
2.000		3.52	611

Figure 7: Education level statistics for parents.

Parent Education Level

1-no high school

2-some high school

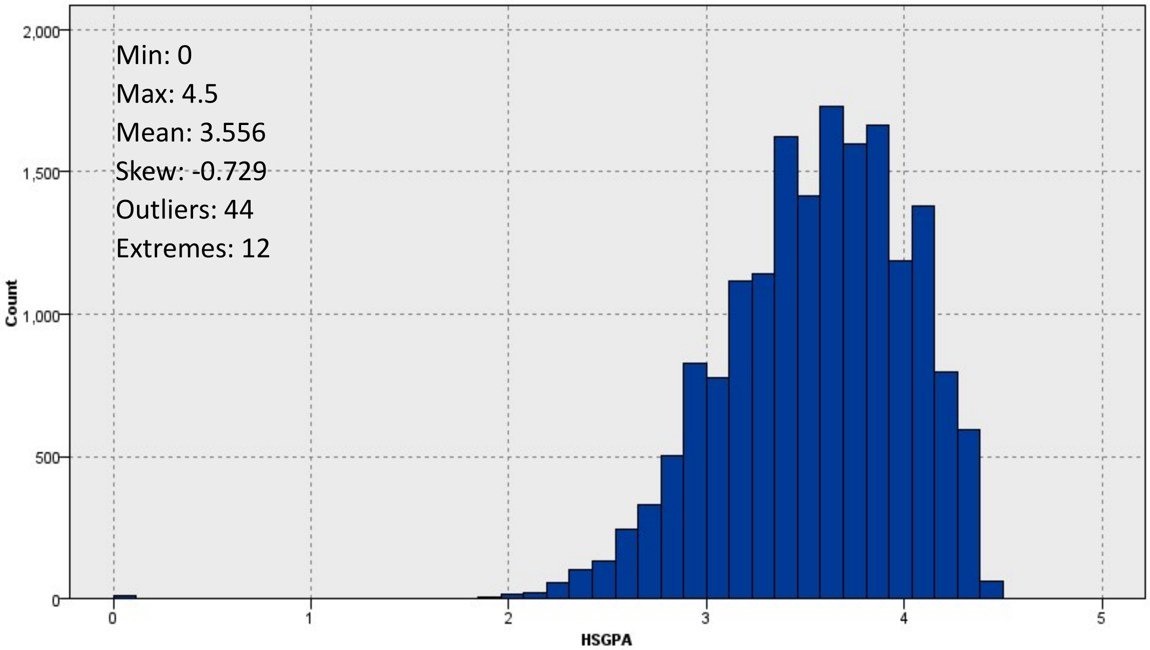
3-high school graduate

4-some college

5-associate degree graduate

6-four-year college degree graduate

7-postgraduate



Min: 0
Max: 4.5
Mean: 3.556
Skew: -0.729
Outliers: 44
Extremes: 12

Figure 8: Histogram for GPA.

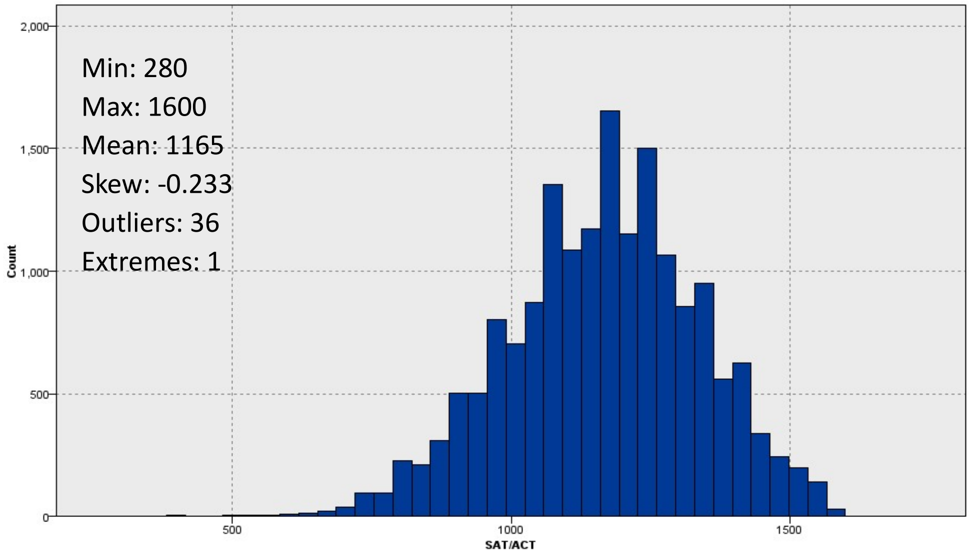


Figure 9: Histogram for SAT/ACT.

<u>Part 3 – Data Preparation:</u>

As a practical guideline for data preparation and modeling we will follow the general outline below:

- **Standardization of data**. Since we will consider using Principle Component Analysis (PCA) to address issues with multicollinearity in later steps, standardization will play an important role. Our approach will use z-scoring. Z-score standardization is generally preferred for its robustness, interpretability, and compatibility with the covariance-based approach of PCA.
- **Feature Selection**. We include a node that will filter the features unnecessary to the analysis. This was done early on where we eliminated "Applicant," "Enrolled," and "College_GPA."
- **Stepwise Regression.** This is a process in SPSS where remaining inputs are used to develop the best model based on the capabilities of the software. The last model will contain inputs most relevant to the analysis.
- **Input interactions**. From the remaining inputs after stepwise, we take these and study them for interactions. We will do this manually in SPSS based observations from correlation studies and our knowledge of the domain of the data.
- **Stepwise Regression**. We re-run the study to include interaction variables.
- **Enforce inputs**. We enforce inputs that need to be there because of the interactions based on model performance measures such as model and input significance, Hosmer-Lemeshow Tests, analysis of correlations, VIF values, coincidence matrices, checks for model over or underfit based on data partitioning, and others as required.
- **PCA**. We run PCA using continuous inputs and interaction terms that have at least one continuous input in them. PCA will help mitigate issues with multicollinearity.
- **Model Audit**. Intermittently, we will be converting the target binary variable to continuous and run Multiple Linear Regression to ensure VIF values are not outside of acceptable ranges.

Detailed steps for data preparation:

Step 1: In IBM SPSS Modeler, we imported the excel file "College Admissions."

- We previewed the data and assessed the condition. As noted, we found one case, Applicant 17310, had a typo graphical error, "Math & ScienceMath & Science &. We edited the entry to resolve the error. We found no empty or null cells.

Step 2: We set flag variables for the binary and categorical types. We standardized inputs using z-scoring to make the data amenable to PCA.

- The "Restructure" node in SPSS was used to restructure the inputs into Flag variables, that is 1 and 0. Fields restructured for the analysis are as follows:
    - Gender: "Gender_F," female = 1, otherwise 0.
    - Admitted: "Admitted_Yes" = 1, otherwise 0.
    - College:
        - "College_Business & Economics" = 1, otherwise 0.
        - "College_Math & Science" = 1, otherwise 0.
        - If both are 0 then "Arts and Letters" is the default.
- We used the "Set Globals" node to standardize continuous inputs.
    - Four fields were selected for standardization:
        - "Edu_Parent1"
        - "Edu_Parent2"
            - These fields are based on the college education range 1 – 7 (**Figure 7**).
        - "HSGPA"
        - "SAT/ACT"
- We used the filter node to filter unnecessary fields once the flag variables were established. Final fields are as follows with the new input name and datatype:
    - For White
        - New input: Data type = Flag, White = 1, otherwise 0.
    - For Asian
        - New input: Data type = Flag, Asian = 1, otherwise 0.
    - For Gender
        - New input: Data type = Flag, Gender_F (Female) = 1, otherwise 0.
    - For Admitted
        - New input: Data type = Flag, Admitted_Yes = 1, otherwise 0.
    - For College
        - New input: Data type = Flag, College_Business & Economics = 1, otherwise 0.
        - New input. Data type = Flag, College_Math & Science = 1, otherwise 0.
        - The default is College_Arts&Letters.
    - For Edu_Parent1:
        - New input: Data type = Continuous, Edu_Parent1_Z.
    - For Edu_Parent2:
        - New input: Data type = Continuous, Edu_Parent2_Z.
    - For HSGPA
        - New input: Data type = Continuous, HSGPA_Z.

- SAT/ACT
  - New input: Data type = Continuous, SAT/ACT_Z.

Step 3: Remove outliers and extremes.

- We created a table of our revised dataset and exported it after removing outliers and extremes based on Interquartile ranges from upper and lower quartiles. Outliers > 1.5 and Extremes > 3. The number of cases in the dataset was reduced to 17,172 cases from 17,339. This helped mitigate skew issues in SAT/ACT_Z and HSGPA_Z.
- For SAT/ACT, this dropped skew from -0.233 to -0.110. The skew was nearly cut in half (**Figure 10**).
- For HSGPA, the skew was reduced from -0.729 to -0.369 (**Figure 11**).

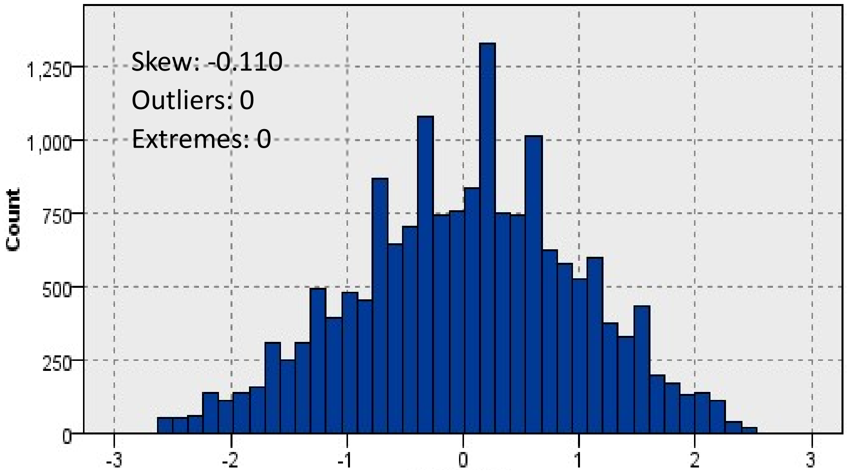


Figure 10: Skew is cut in half after removing outliers and extremes for SAT/ACT Z.

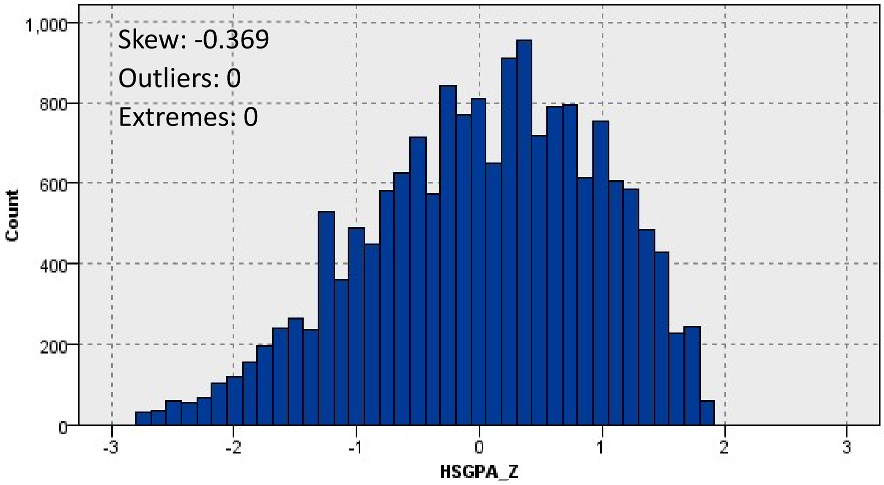
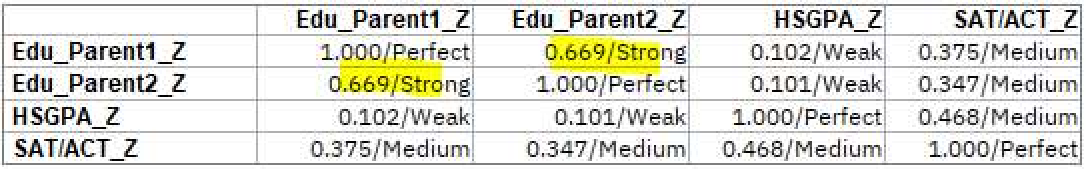


Figure 11: Skew is reduced by almost half for HSGPA_Z.

Part 4 – Modeling:

Step 1: Explore interactions.

- Pearson correlations were run in SPSS to check for multicollinearity issues. The fields Edu_Parent1_Z and Edu_Parent2_Z were found to be highly correlated (**Figure 12**). Effects of multicollinearity can be managed by PCA, dropping an input, or introducing interaction variables. Other correlations were found to be weak or medium.

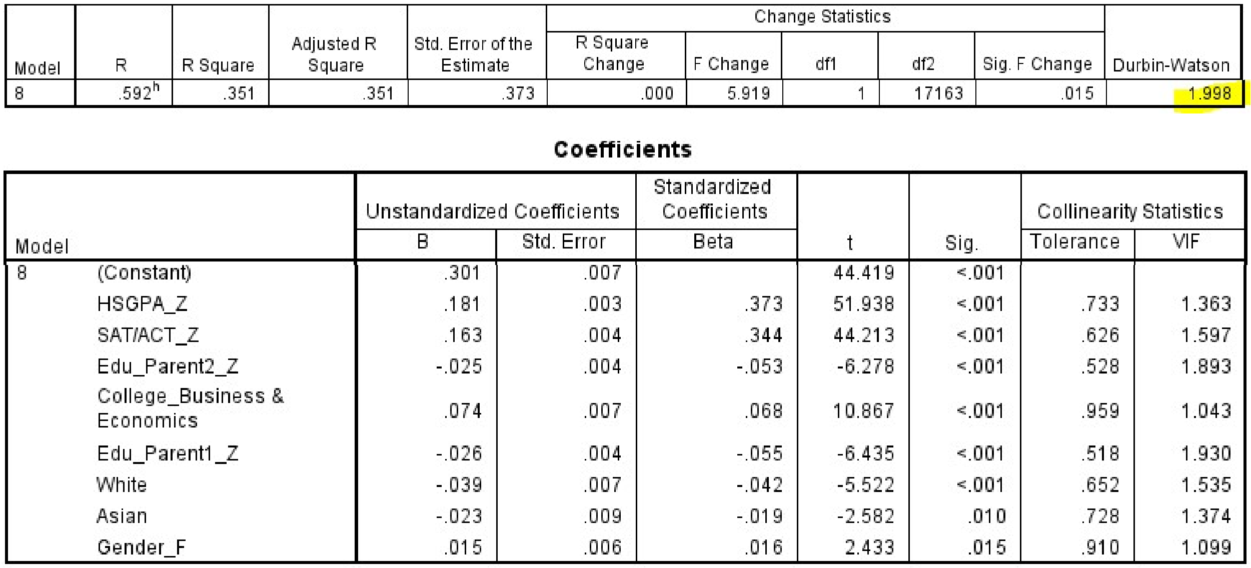


	Edu_Parent1_Z	Edu_Parent2_Z	HSGPA_Z	SAT/ACT_Z
Edu_Parent1_Z	1.000/Perfect	0.669/Strong	0.102/Weak	0.375/Medium
Edu_Parent2_Z	0.669/Strong	1.000/Perfect	0.101/Weak	0.347/Medium
HSGPA_Z	0.102/Weak	0.101/Weak	1.000/Perfect	0.468/Medium
SAT/ACT_Z	0.375/Medium	0.347/Medium	0.468/Medium	1.000/Perfect

Figure 12: Pearson correlations.

Step 2: Preliminary Modeling and Feature Selection

- To facilitate feature selection, we converted the target "Admitted_Yes" to a continuous datatype to analyze the data in a Multiple Linear Regression (MLR) scenario. We looked at input and model significance, variance inflation factors (VIF), adjusted R^2, Durban-Watson values, and other measures using stepwise in SPSS. The software stopped processing at step 8. The only input excluded was "College_Math & Science." All inputs were found to be significant; the Durban-Watson value was acceptable, VIF values were acceptable, the model was found to be significant, however adjusted R^2 is poor (**Figures 13 and 14**).

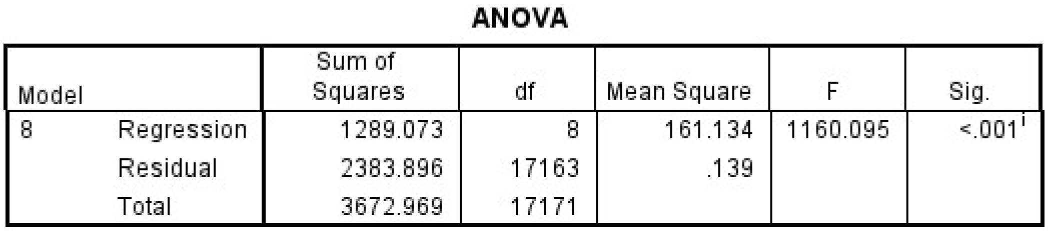


					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
8	.592^h	.351	.351	.373	.000	5.919	1	17163	.015	1.998

Coefficients

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
8	(Constant)	.301	.007		44.419	<.001		
	HSGPA_Z	.181	.003	.373	51.938	<.001	.733	1.363
	SAT/ACT_Z	.163	.004	.344	44.213	<.001	.626	1.597
	Edu_Parent2_Z	-.025	.004	-.053	-6.278	<.001	.528	1.893
	College_Business & Economics	.074	.007	.068	10.867	<.001	.959	1.043
	Edu_Parent1_Z	-.026	.004	-.055	-6.435	<.001	.518	1.930
	White	-.039	.007	-.042	-5.522	<.001	.652	1.535
	Asian	-.023	.009	-.019	-2.582	.010	.728	1.374
	Gender_F	.015	.006	.016	2.433	.015	.910	1.099

Figure 13: Multiple linear regression. Target value "Admitted_Yes" converted to continuous to evaluate VIF and other values.

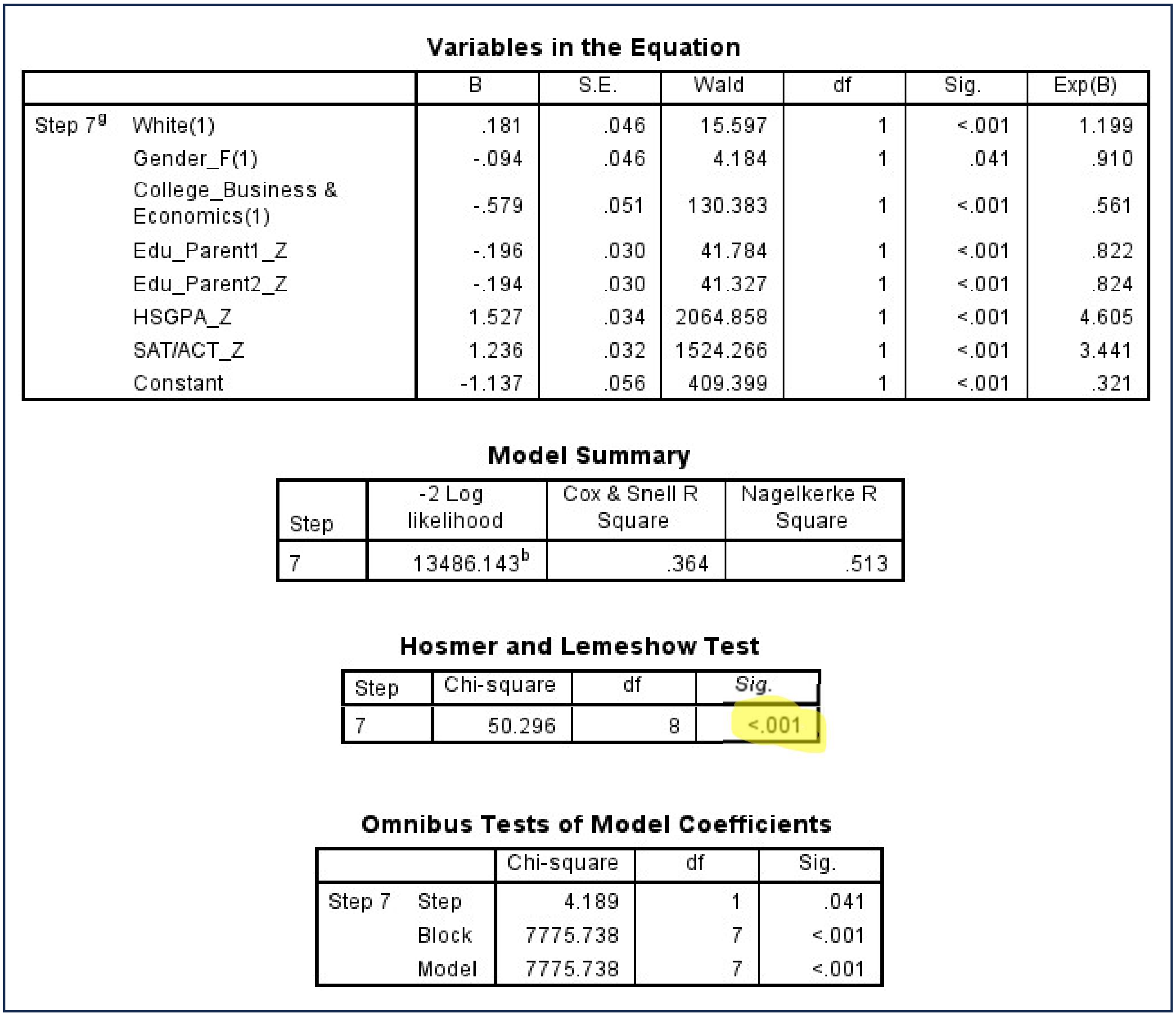


ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
8	Regression	1289.073	8	161.134	1160.095	<.001[i]
	Residual	2383.896	17163	.139		
	Total	3672.969	17171			

Figure 14: MLR, Model is significant.

In a similar fashion, we ran the model in a logistic regression (LR), again using stepwise in SPSS. The model stopped at the 7th modeling step (**Figure 15**). This did not include the input "Asian" in the model as was in found in the MLR.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 7[g]	White(1)	.181	.046	15.597	1	<.001	1.199
	Gender_F(1)	-.094	.046	4.184	1	.041	.910
	College_Business & Economics(1)	-.579	.051	130.383	1	<.001	.561
	Edu_Parent1_Z	-.196	.030	41.784	1	<.001	.822
	Edu_Parent2_Z	-.194	.030	41.327	1	<.001	.824
	HSGPA_Z	1.527	.034	2064.858	1	<.001	4.605
	SAT/ACT_Z	1.236	.032	1524.266	1	<.001	3.441
	Constant	-1.137	.056	409.399	1	<.001	.321

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
7	13486.143[b]	.364	.513

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
7	50.296	8	<.001

Omnibus Tests of Model Coefficients

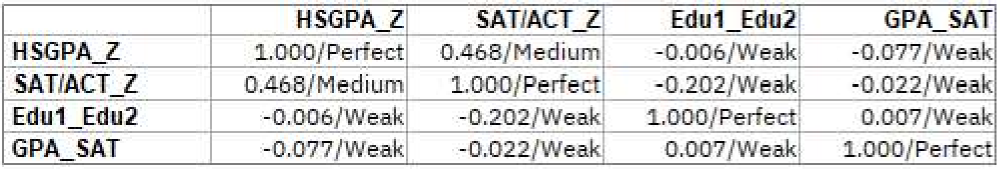
		Chi-square	df	Sig.
Step 7	Step	4.189	1	.041
	Block	7775.738	7	<.001
	Model	7775.738	7	<.001

Figure 15: Logistic regression outputs for input and model significance and various model metrics.

From observations made int **Figure 15**, that the model is significant, inputs are significant, Nagelkerke R^2 is low, but not objectionable, however the Hosmer and Lemeshow Test for Significance is <0.001. The Hosmer-Lemeshow test examines the hypothesis, H_0 the model fits the data well vs. H_1, the models do not fit the data well. Our threshold level of significance is 0.05, i.e., we would not reject the null hypothesis H_0, if “Sig.” is > 0.05. The model significance < 0.001. This suggests that the data is not fit by the model – we reject H_0. We found it necessary to explore interaction between inputs to resolve the model fit issue.

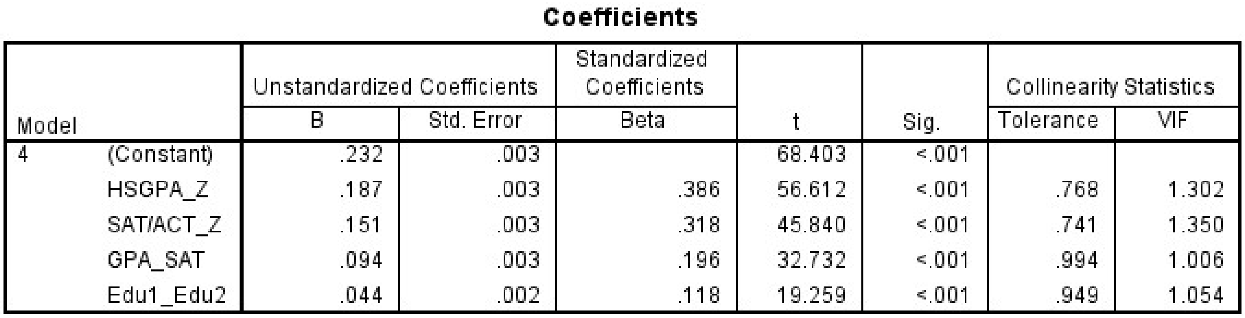
Step 5: Develop interaction variables:

We suspected an issue with multicollinearity within the inputs “Edu_Parent1” and “Edu_Parent2” from the earlier analysis of collinearity (**Figure 2**). We began to analyze interactions beginning with “Edu_Parent1” and “Edu_Parent2,” “HSGPA_Z and SAT/ACT_Z,” i.e., the continuous variable types in our model. Interaction variables “Edu1_Edu2” and GPA_SAT were developed comprised of the related inputs multiplied together. The Pearson correlation examined for these new inputs, **Figure 16**, was found to mitigate strong correlations. As well, we examined VIF values in a MLR scenario again and found VIF values were acceptable (**Figure 17**).



	HSGPA_Z	SAT/ACT_Z	Edu1_Edu2	GPA_SAT
HSGPA_Z	1.000/Perfect	0.468/Medium	-0.006/Weak	-0.077/Weak
SAT/ACT_Z	0.468/Medium	1.000/Perfect	-0.202/Weak	-0.022/Weak
Edu1_Edu2	-0.006/Weak	-0.202/Weak	1.000/Perfect	0.007/Weak
GPA_SAT	-0.077/Weak	-0.022/Weak	0.007/Weak	1.000/Perfect

Figure 16: Pearson correlation including the new variable "Edu1_Edu2."

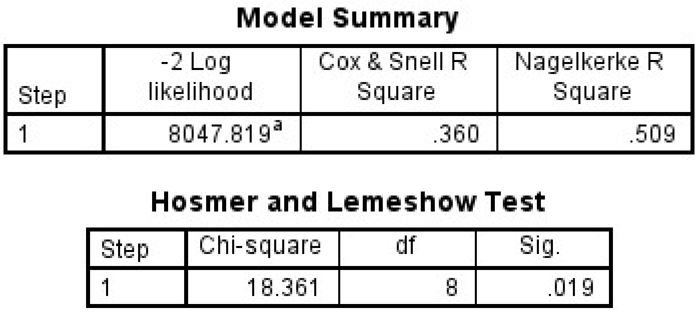


Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
4	(Constant)	.232	.003		68.403	<.001		
	HSGPA_Z	.187	.003	.386	56.612	<.001	.768	1.302
	SAT/ACT_Z	.151	.003	.318	45.840	<.001	.741	1.350
	GPA_SAT	.094	.003	.196	32.732	<.001	.994	1.006
	Edu1_Edu2	.044	.002	.118	19.259	<.001	.949	1.054

Figure 17: MLR with new input "Edu1_Edu2" introduced.

With the interactions between the inputs Edu_Parent1_Z and Edu_Parent2_Z considered; we re-ran the Logistic Regression model with, again, only the continuous inputs. The Hosmer-Lemeshow significance value did improve, but we were looking for a threshold of > 0.05 significance level, further the Nagelkerke R^2 value decreased (**Figure 18**).



Model Summary

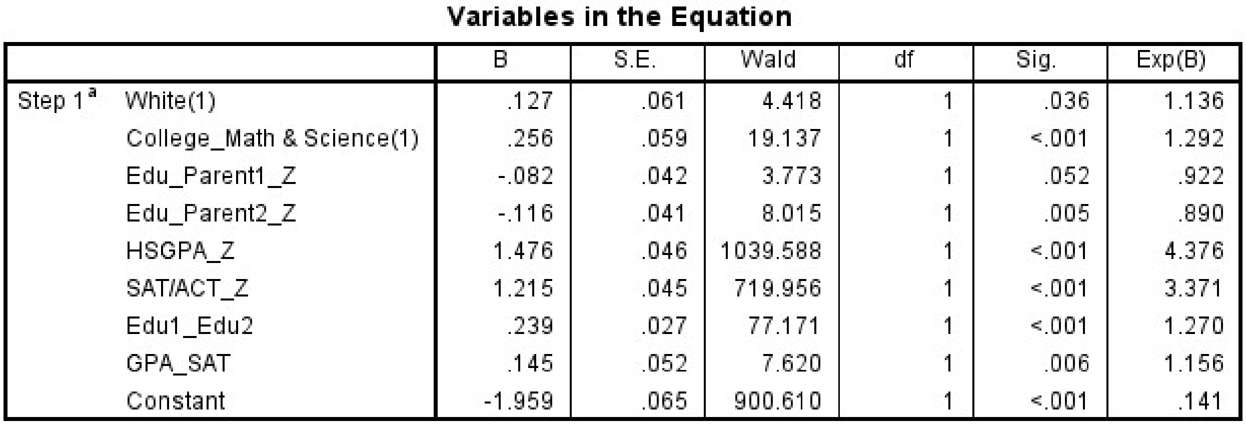
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	8047.819[a]	.360	.509

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	18.361	8	.019

Figure 18: Logistic regression with continuous inputs and the interaction vvariable Edu1_Edu2. Hosmer Lemeshow improved, but Nagelkerke R2 decreased.

In an iterative fashion, we explored multiple interactions between inputs, including flag and continuous variables, utilizing PCA where we found VIF values were an issue, in a one-by-one fashion. We continued to check Hosmer and Lemeshow Test values, Nagelkerke R^2, and utilized Pearson correlations to investigate issues with multicollinearity during the model development process. The set of inputs we found to produce the most robust model, **Model 3**, are found in **Figure 19**. Note, we found factors derived from PCA were not useful in developing a stronger model.

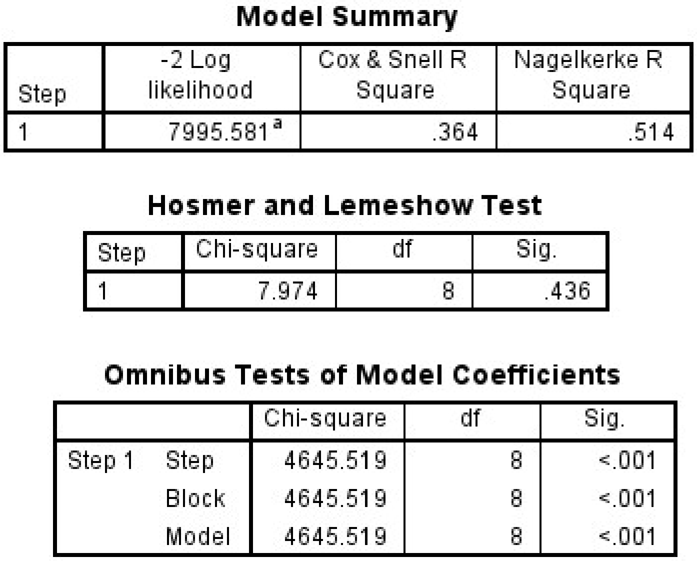


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	White(1)	.127	.061	4.418	1	.036	1.136
	College_Math & Science(1)	.256	.059	19.137	1	<.001	1.292
	Edu_Parent1_Z	-.082	.042	3.773	1	.052	.922
	Edu_Parent2_Z	-.116	.041	8.015	1	.005	.890
	HSGPA_Z	1.476	.046	1039.588	1	<.001	4.376
	SAT/ACT_Z	1.215	.045	719.956	1	<.001	3.371
	Edu1_Edu2	.239	.027	77.171	1	<.001	1.270
	GPA_SAT	.145	.052	7.620	1	.006	1.156
	Constant	-1.959	.065	900.610	1	<.001	.141

Figure 19: Final model inputs, Model 3.

Summary statistics are provided in **Figure 20**. Note, we cannot reject the model based on the Hosmer – Lemmeshow output. Nagelkerke R^2 is not strong, but acceptable. The model is significant, and all inputs are significant with the exception of Edu_Parent1_Z. However, we are required to include individual inputs for those that are multiplied together in order to complete the model. A summary table of models evaluated is provided in **Figure 21**.



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	7995.581[a]	.364	.514

Hosmer and Lemeshow Test

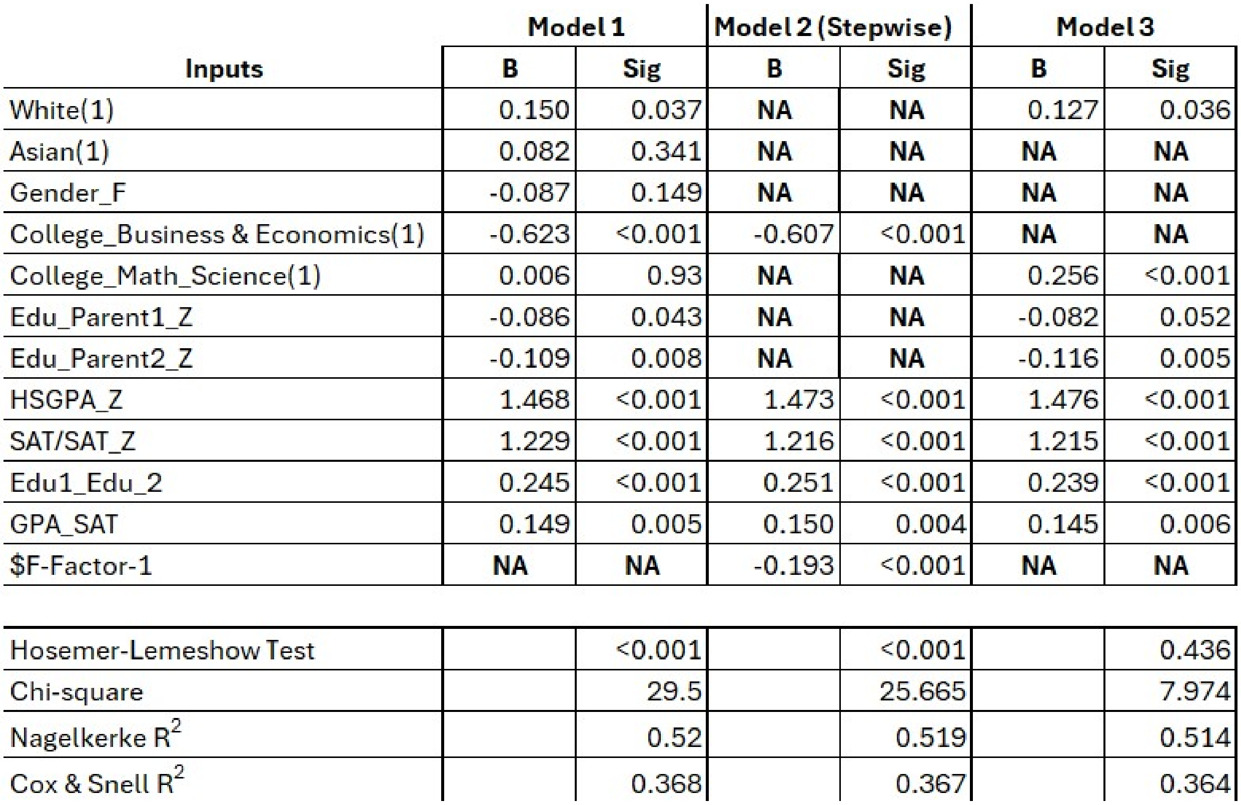
Step	Chi-square	df	Sig.
1	7.974	8	.436

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	4645.519	8	<.001
	Block	4645.519	8	<.001
	Model	4645.519	8	<.001

Figure 20: Model is significant. Nagelkerke R2 is acceptable, and Hosmer and Lemeshow Test confirms model veracity.

Provided, is a summary of 3 models evaluated in **Figure 21**. Model 1 is inclusive of all inputs with the exception of one factor derived from PCA, $F-Factor-1. We used stepwise in SPSS using the same inputs to derive an additional model, Model 2, that did include the PCA derived input. In both of these cases, the models fail the Hosmer – Lemmeshow test. We then took an iterative approach to evaluate inputs one at a time to develop a stronger model, Model 3.



	Model 1		Model 2 (Stepwise)		Model 3	
Inputs	**B**	**Sig**	**B**	**Sig**	**B**	**Sig**
White(1)	0.150	0.037	**NA**	**NA**	0.127	0.036
Asian(1)	0.082	0.341	**NA**	**NA**	**NA**	**NA**
Gender_F	-0.087	0.149	**NA**	**NA**	**NA**	**NA**
College_Business & Economics(1)	-0.623	<0.001	-0.607	<0.001	**NA**	**NA**
College_Math_Science(1)	0.006	0.93	**NA**	**NA**	0.256	<0.001
Edu_Parent1_Z	-0.086	0.043	**NA**	**NA**	-0.082	0.052
Edu_Parent2_Z	-0.109	0.008	**NA**	**NA**	-0.116	0.005
HSGPA_Z	1.468	<0.001	1.473	<0.001	1.476	<0.001
SAT/SAT_Z	1.229	<0.001	1.216	<0.001	1.215	<0.001
Edu1_Edu_2	0.245	<0.001	0.251	<0.001	0.239	<0.001
GPA_SAT	0.149	0.005	0.150	0.004	0.145	0.006
$F-Factor-1	**NA**	**NA**	-0.193	<0.001	**NA**	**NA**
Hosemer-Lemeshow Test		<0.001		<0.001		0.436
Chi-square		29.5		25.665		7.974
Nagelkerke R^2		0.52		0.519		0.514
Cox & Snell R^2		0.368		0.367		0.364

Figure 21: Comparing models.

Part 5 – Findings and Evaluation:

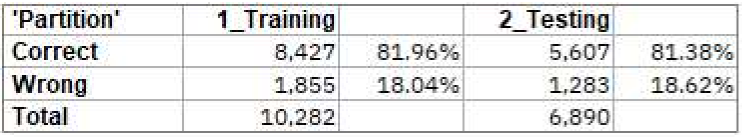
A rigorous evaluation of the inputs relative to the most appropriate model was conducted. After data preparation, we investigated all inputs initially. We used SPSS stepwise to facilitate the best inputs to use, then built upon this work and applied PCA methods where multicollinearity was deemed to be a problem.

Model Reviews:

Model 1: This was an initial review of the inputs in SPSS. We found the input "Asian" was not significant in the model. The Nagelkerke R^2 value was acceptable, however the model was deficient relative to the Hosemer-Lemeshow Tests. Further its Chi-Square value was the poorest among all models.

Model 2: Model 2 was developed after we ran stepwise in SPSS. Several inputs were dropped from the model. The remaining inputs were all significant. However, as in the case of Model 1, the Hosemer-Lemeshow Tests score was not acceptable. Its Chi-Square value improved but was poor relative to Model 3.

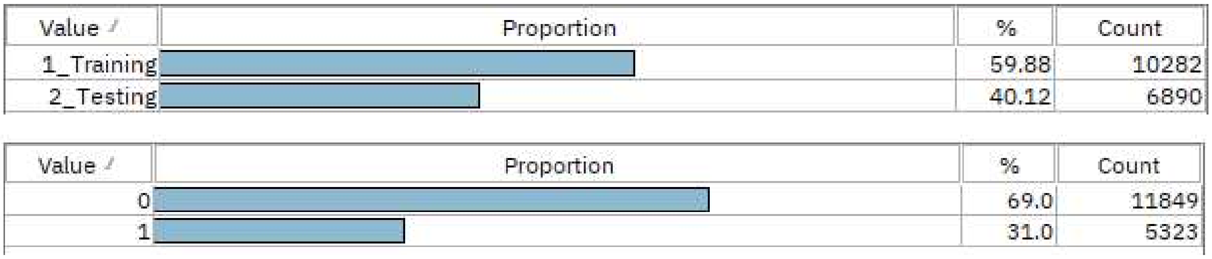
Model 3: Due to issues with Models' 1 and 2 performances relative to key statistical metrics, i.e., Hosemer-Lemeshow Tests as well as the Chi-Square test, we find this model is the most favorable to predict Admission (**Figure 21**). The Hosemer-Lemeshow Tests score is strong, strongly suggesting we do not reject the model (H_0). Further, It's Chi-Square value is the best among the three. The model is significant as well as the inputs – however, the input "Edu_Parent1_Z" is slightly over our threshold for rejection (0.05). We are obligated to it keep in the model as is it part of the interaction variable "Edu1_Edu2." As a final check for Model 3, we evaluated the comparison, Matrix. Based on a 60/40 split for train/test of the data, we find there is no overfit or underfit (**Figure 22**). All in all, the model's performance relative to these metrics is strong.



'Partition'	1_Training		2_Testing	
Correct	8,427	81.96%	5,607	81.38%
Wrong	1,855	18.04%	1,283	18.62%
Total	10,282		6,890	

Figure 22: Comparison matrix for test. vs. train.

We checked to confirm we have enough cases for the analysis based on the number of inputs involved. The rule of thumb for the sample size for Logistic Regression is a minimum of 10 cases for each independent variable or input multiplied by the number of variables or inputs. Divide that number by the least probability of the least frequent outcome. Then ensure we have at least that many of cases in the test dataset. There are 8 input variables in Model 3. To help us calculate the minimum sample size, we added a graph to the Partition node in SPSS to find the number of cases in the training and test set (**Figure 23**). Note that, despite our setting of 60% train and 40% test split, SPSS makes the most appropriate split based on the data. Then we plotted the proportionality plot for the number of ones and zeros in the entire dataset for the target "Admission_Yes." The least occurring were the 1's; 31%.



Value	Proportion	%	Count
1_Training		59.88	10282
2_Testing		40.12	6890

Value	Proportion	%	Count
0		69.0	11849
1		31.0	5323

Figure 23: Proportionality plots for training and test data; counts for ones and zeros in the target "Admission_Yes."

$$number\ of\ inputs = \frac{10\ x\ 8}{0.31} = 258$$

A minimum of 258 cases are required. We have 10,282. The number of cases is sufficient for the model.

Part 6 – Deployment and Recommendations:

Our recommendation for a model to predict the target "Admission" then, is **Model 3**. The next step we take is to fully define the estimated model for predicting the target.

Given the inputs, we developed the estimated logistic model using Admission as the target. The estimated model is as follows:

$$\hat{\pi}(x) = \frac{e^{\left\{\begin{matrix}-1.959+.145*GPASAT+.239*Edu1Edu2+1.215*SATACT+1.476*HSGPA \quad (-.116)*EduParent2 \\ (-.082)*EduParent1Z+.256*College\_Math\ \ Science+.127*White\end{matrix}\right\}}}{1+\ e^{\left\{\begin{matrix}-1.959+.145*GPASAT+.239*Edu1Edu2+\ .215*SATACT+1.476*HSGPAZ+(-.116)*EduParent2Z+ \\ (-.082)*EduParent1Z+.256*College\_Math\ \ Science+.127*White\end{matrix}\right\}}}$$

The maximum likelihood estimates for the betas are:

$b_0 = -1.959$, for the constant.

$b_1 = 0.145$, for GPA_SAT.

$b_2 = 0.239$, Edu1_Edu2

$b_3 = 1.215$, for SAT_ACT_Z.

$b_4 = 1.476$, for HSGPA_Z

$b_5 = -0.116$, for Edu_Parent2_Z

$b_6 = -0.082$, for Edu_Parent1_Z

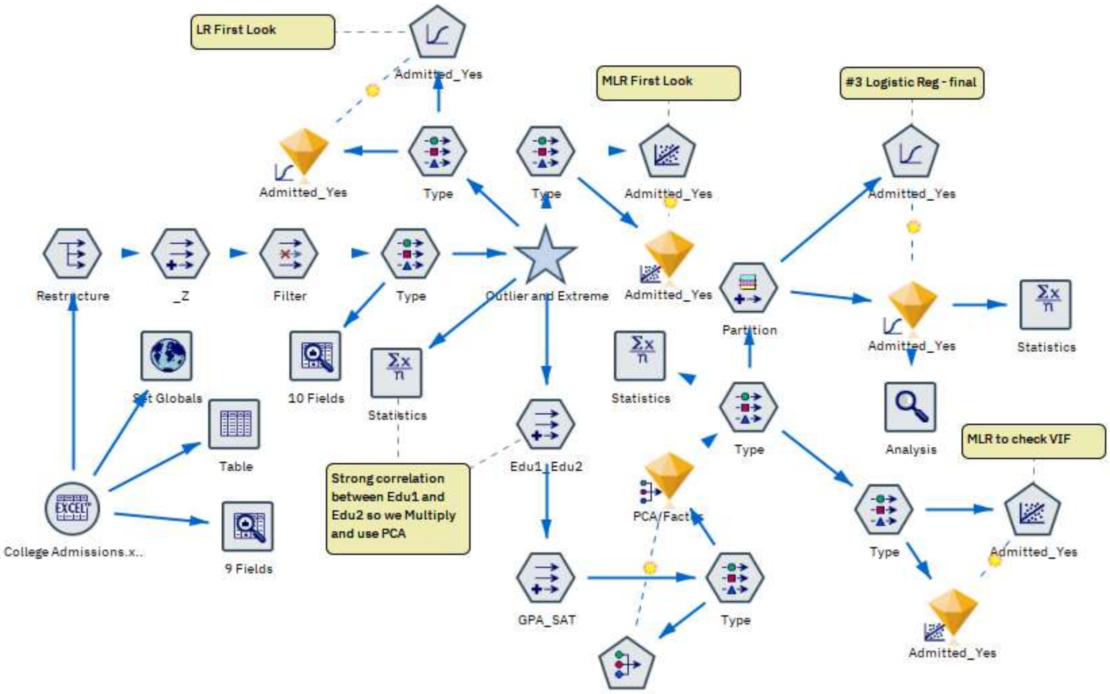
$b_7 = 0.256$, for College_Math & Science

$b_8 = 0.127$, for White

The related odds ratios are discussed as follows:

> GPA_SAT = 1.156. For each unit increase in GPA * SAT there is a 15.6% increase in the odds for admission, other inputs held constant.
>
> Edu1_Edu2 = 1.270. For each unit increase in Edu1 * Edu2 there is a 27% increase in the odds for admission, other inputs held constant.
>
> SAT/ACT_Z = 3.371. For each unit increase in SAT/ACT_Z, there is a 3.371 times increase in the odds for admission, other inputs held constant.
>
> HSGPA_Z = 4.376 For each unit increase in HSGPA_Z there is a 4.376 time increase in odds for admission, other inputs held constant.
>
> Edu_Parent2_Z = 0.890. (1-0.890 = 0.11) For each unit increase in Edu_Parent2, there is an 11% decrease in on odds of admission other inputs held constant.
>
> Edu_Parent1_Z = 0.922. (1-0.922 = 0.078). For each unit increase in Edu_Parent1_Z, there is a 7.8% decrease in odds of admission other inputs held constant.
>
> College_Math & Science = 1.292. For each unit increase in College_Math & Science, the odds of admission increases 29.92% other inputs held constant.
>
> White = 1.136. For each unit increase in White, the odds of Admission increases 13.6% other inputs held constant.

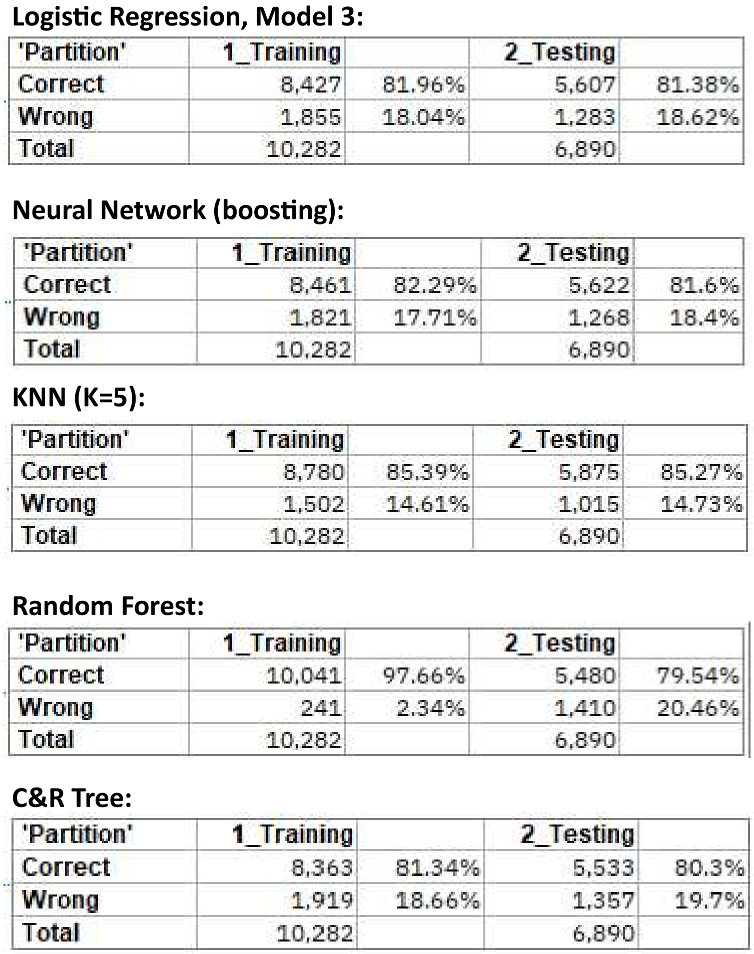
The final model stream developed is provided in **Figure 24**.



Strong correlation between Edu1 and Edu2 so we Multiply and use PCA

Figure 24: Recommended SPSS Model 3 stream.

To verify our recommendation, we compared Model 3 developed with Logistic Regression to Neural Network, Random Trees, KNN, Random Forest, and the C&R Tree models. To compare the approaches, we looked at Coincidence Matrices using the Analysis node in SPSS, with train/test data at a 60/40 split (**Figure 25**).



Logistic Regression, Model 3:

'Partition'	1_Training		2_Testing	
Correct	8,427	81.96%	5,607	81.38%
Wrong	1,855	18.04%	1,283	18.62%
Total	10,282		6,890	

Neural Network (boosting):

'Partition'	1_Training		2_Testing	
Correct	8,461	82.29%	5,622	81.6%
Wrong	1,821	17.71%	1,268	18.4%
Total	10,282		6,890	

KNN (K=5):

'Partition'	1_Training		2_Testing	
Correct	8,780	85.39%	5,875	85.27%
Wrong	1,502	14.61%	1,015	14.73%
Total	10,282		6,890	

Random Forest:

'Partition'	1_Training		2_Testing	
Correct	10,041	97.66%	5,480	79.54%
Wrong	241	2.34%	1,410	20.46%
Total	10,282		6,890	

C&R Tree:

'Partition'	1_Training		2_Testing	
Correct	8,363	81.34%	5,533	80.3%
Wrong	1,919	18.66%	1,357	19.7%
Total	10,282		6,890	

Figure 25: Model comparing, logistic regression, neural network, random trees, random forest, and C&R Tree.

Model Accuracies:

Refer to **Figure 26** for a summary of model accuracies.

Model Discussion:

Logistic Regression, Model 3: Model 3 logistic regression was found to be the most pertinent among the logistic regression models. We used Stepwise in SPSS to assist in the selection of features then modified inputs to get the best output based on experimentation. We vetted the models with the metrics noted in **Figure 21** to conclude this model was the most relevant among the logistic group.

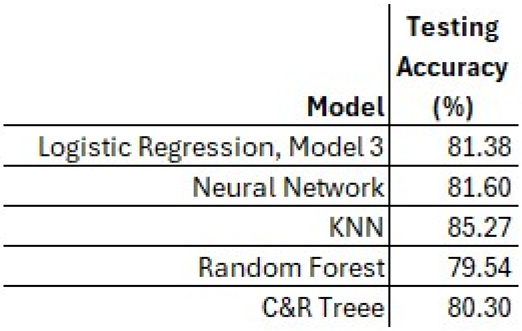
Neural Network: We experimented with the options Multilayer Perceptron (MLP), Radial Basis Function, Boosting, Bagging, and a standard model. We let SPSS automatically compute the number of units for the hidden layers. We found the optimal model accuracy was using MLP and Boosting. All inputs were considered. Accuracy was close to our Logistic model.

KNN (K=5): We set K=5 for the K-Nearest Neighbor model. We let SPSS perform feature selection based select inputs, namely continuous inputs for the two parent types, the interaction between them, high school GPA, and the interaction between GPA and SAT. This model proved to be the most accurate among the models.

Random Forest: With number of trees to build = 5, leaf node minimum = 1, Number of Features to use for splitting set to “Auto,” and using the bootstrap option for Model Building, accuracy was the poorest of the models. The model severely overfit the data as well.

C&RT: For the C&RT model, we experimented with inputs, and kept the main objective to “Build a single tree.” Even though we attempted to boost the model accuracy, it was keeping the main objective at the single tree option that produced the most accurate model. However, at 80.03% it was not as accurate as the Logistic Model.

Our conclusion among the models is the K-Nearest Neighbor with K set at 5 neighbors was the most powerful. The accuracy was best; there was almost no overfitting at all. A screenshot of the SPSS stream for these models is provided in **Figure 27**.



Model	Testing Accuracy (%)
Logistic Regression, Model 3	81.38
Neural Network	81.60
KNN	85.27
Random Forest	79.54
C&R Treee	80.30

Figure 26: Testing accuracy summary.

We took an additional step and ran a Logistic Regression in the software model “R.” Here, we selected various inputs and input interactions to develop an accurate Logistic model. Accuracy was very close to the model developed in SPSS at about 82.2% with almost no overfitting (**Figure 28**).

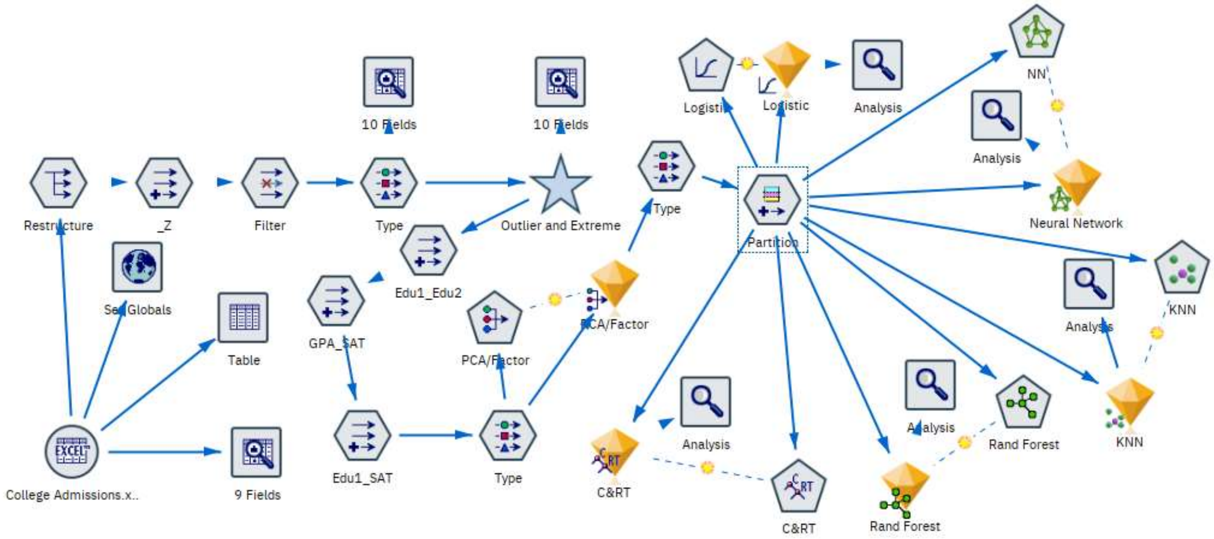
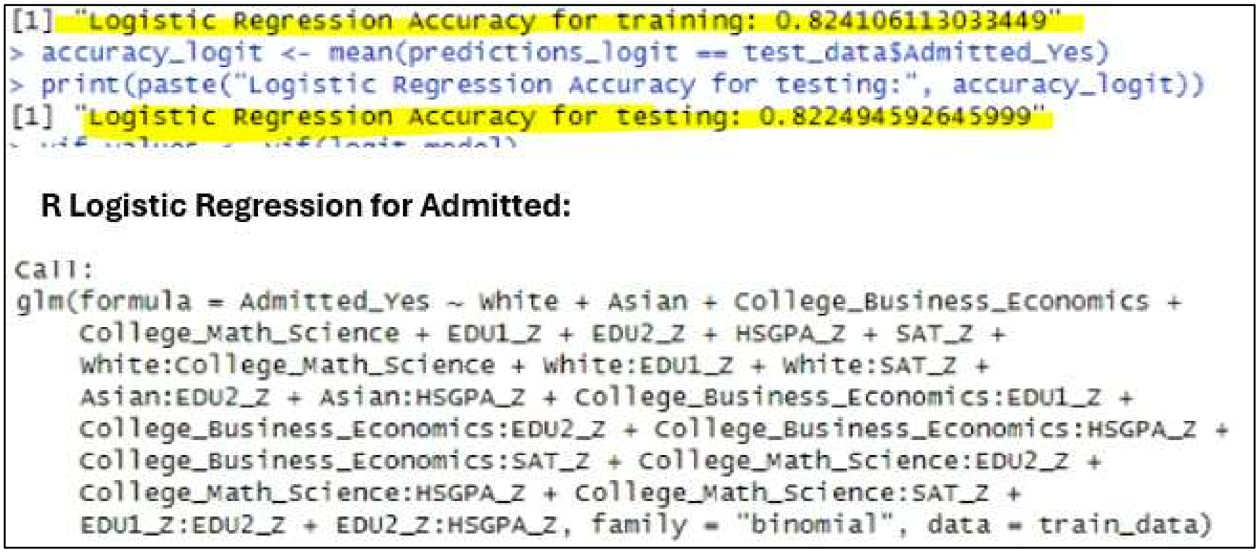


Figure 27: Model stream for Logistic, Neural Net, KNN, and C&RT.



```
[1] "Logistic Regression Accuracy for training: 0.824106113033449"
> accuracy_logit <- mean(predictions_logit == test_data$Admitted_Yes)
> print(paste("Logistic Regression Accuracy for testing:", accuracy_logit))
[1] "Logistic Regression Accuracy for testing: 0.822494592645999"
> vif_values <- vif(logit_model))
```

R Logistic Regression for Admitted:

```
Call:
glm(formula = Admitted_Yes ~ White + Asian + College_Business_Economics +
    College_Math_Science + EDU1_Z + EDU2_Z + HSGPA_Z + SAT_Z +
    White:College_Math_Science + White:EDU1_Z + White:SAT_Z +
    Asian:EDU2_Z + Asian:HSGPA_Z + College_Business_Economics:EDU1_Z +
    College_Business_Economics:EDU2_Z + College_Business_Economics:HSGPA_Z +
    College_Business_Economics:SAT_Z + College_Math_Science:EDU2_Z +
    College_Math_Science:HSGPA_Z + College_Math_Science:SAT_Z +
    EDU1_Z:EDU2_Z + EDU2_Z:HSGPA_Z, family = "binomial", data = train_data)
```

Figure 28: R output for Logistic Regression. Comparable to SPSS output for same modeling method.