Predictive Analysis On Student Dropout Data Report

A predictive analysis was carried out on the Student Dropout Data set where predictions were made to see the chances that students would either graduate or drop out of school using a set of variables in the data set. This report has been created to outline the methods used during the analysis and evaluation of the results.

Methods

The first thing which was carried out was the analysis of the dataset. The data file was known as Dropout_data and was part of the Student Dropout Data set. The following variables were included: Moqual (Mother's Qualification), Faqual (Father's Qualification), Admingrade (Admission Grade), Eduspecneeds (Educational special needs), Prevgrade (Previous qualification (grade)), Ownschshp (Scholarship holder), Prevqual (Previous qualification), Debtor (Debtor), Paidfeetodate (Tuition fees up to date), Gender (Gender), Secsemgrade (Curricular units 2nd sem (grade)), Target (Graduate or dropout), Course (Course of study), Displaced (Changed Schools Before).

The libraries pandas, numpy, sklearn, and statsmodels were used in this predictive analysis project because these libraries are known to provide the required functionality for handling data, building and testing models, and performing statistical analysis. Throughout the project, the main analytical tool was a multivariate linear regression model because it allows the modeling between a dependent variable and multiple independent variables. It shows the effects of multiple independent variables on the dependent variable, and can be used to make predictions for new data based on the values of the independent variables.

After importing the required libraries and the data file, the first step which was carried out was creating a multivariate linear regression model with all the variables to see the relationship between the dependent variable (target) and the multiple independent variables (the remaining variables). The next step was to find the model's accuracy. The code fits a logistic regression model to the test data and predicts the probability of a target event occurring based on input

variables. It then creates new columns for the predicted binary values and the accuracy of the model, and prints out the number of true and false predictions, as well as the prediction accuracy.

Generalized Linear Model Regression Results

Dep. Variable:	target	No. Observations:	2904
Model:	GLM	Df Residuals:	2890
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-974.04
Date:	Fri, 24 Feb 2023	Deviance:	1948.1
Time:	17:30:48	Pearson chi2:	4.32e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.9898	0.772	-5.168	0.000	-5.503	-2.477
moqual	-0.0106	0.004	-2.384	0.017	-0.019	-0.002
faqual	0.0039	0.005	0.856	0.392	-0.005	0.013
admingrade	0.0130	0.005	2.525	0.012	0.003	0.023
eduspecneeds	-0.0430	0.507	-0.085	0.933	-1.038	0.952
prevgrade	-0.0031	0.005	-0.566	0.572	-0.014	0.008
ownschshp	1.2680	0.153	8.291	0.000	0.968	1.568
prevqual	-0.0119	0.006	-2.152	0.031	-0.023	-0.001
debtor	-0.8303	0.206	-4.035	0.000	-1.234	-0.427
paidfeetodate	3.0597	0.266	11.510	0.000	2.539	3.581
gender	-0.7536	0.121	-6.239	0.000	-0.990	-0.517
secsemgrade	0.4861	0.033	14.826	0.000	0.422	0.550
course	-0.0005	4.88e-05	-10.545	0.000	-0.001	-0.000
displaced	0.0014	0.120	0.012	0.991	-0.233	0.236

Through the output of the multivariate linear regression model with all the variables (as seen above), it was visible that there were some variables which were not statistically significant such as faqual. So, the next step was creating another multivariate linear regression model with only the variables which were statistically significant (P-value (P> |z|) < 0.05) and the variable course was also dropped as a students course does not really have the probability of predicting graduation or dropout as you have the ability to change your course of study. The next step was to find this model's accuracy using the same steps as the previous model.

Results/Analysis

Accuracy Of Model With All Variables	84.98622%
Accuracy of Model With Dropped Variables	84.02204%

The difference in accuracy between the two variables is not major, hence, it made more sense to pick the model with dropped variables as the more preferred model as it includes only the variables which are statistically significant for the dependent variable (target). The following multivariate linear regression model was produced for this model:

Generalized Linear Model Regression Results

Dep. Variable:	target	No. Observations:	2904
Model:	GLM	Df Residuals:	2896
Model Family:	Binomial	Df Model:	7
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1070.2
Date:	Tue, 21 Feb 2023	Deviance:	2140.4
Time:	06:43:58	Pearson chi2:	2.88e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.3790	0.598	-12.348	0.000	-8.550	-6.208
moqual	-0.0101	0.004	-2.838	0.005	-0.017	-0.003
admingrade	0.0206	0.004	5.283	0.000	0.013	0.028
ownschshp	1.4003	0.148	9.452	0.000	1.110	1.691
debtor	-0.8712	0.197	-4.424	0.000	-1.257	-0.485
paidfeetodate	2.9573	0.246	12.036	0.000	2.476	3.439
gender	-0.7166	0.115	-6.258	0.000	-0.941	-0.492
secsemgrade	0.2743	0.013	21.570	0.000	0.249	0.299

From the coefficients in the model, we can see that the independent variables that are most strongly associated with the prediction are paidfeetodate, ownschshp, and secsemgrade. Students who have paid more fees, have a scholarship, or have a higher grade average are less likely to be at high risk of dropping out. Conversely, students who have outstanding debts, have a lower level of education than their mother, or are female are more likely to be at high risk of dropping out.

Discussion

The model correctly predicted 610 out of 726 observations, resulting in a prediction accuracy of 84.02%. Additionally, there were 116 false predictions, indicating that the model is not 100% accurate in its predictions. Overall, these results suggest that the model is effective in predicting the target variable based on the given independent variables, but there may be some room for improvement in order to increase its accuracy. One thing worth noting is that after removing all variables which were not statistically significant from the first model (with all the variables), in

this model we see that moqual is not statistically significant as it has a P-value of 0.005. This suggests that an increase in moqual is associated with a decrease in the likelihood of the target variable being positive.

Conclusion

In conclusion, the predictive analysis carried out on the Student Dropout Data set was successful in developing a model that predicts the chances of a student either graduating or dropping out of school. The accuracy of the model with dropped variables was found to be 84.02%, and the most significant variables were paidfeetodate, ownschshp, admingrade and secsemgrade. Overall, the model provides valuable insights into the factors that contribute to a student's likelihood of dropping out of school.