**Case Study: College Dropout Prediction Analysis**

Author: Manuel Duran

Bellevue University DSC-630

* **OVERVIEW**

This case study showcases a recent College Dropout Prediction Analysis Project for the American Public University System. The recent college system has seen a staggering number of students beginning to drop out and not finish their degree. The American Public University System collected data to get ahead of the issues and attempt to decrease the number of dropouts that occur year in and year out. For our source data for this project, we have teamed up with several collegiate universities including: Arizona State, Iowa State, Oklahoma State and The University of Alabama. By reviewing the Universities records, the team aspires to discover actionable insights based on analysis.

* **Business Understanding: Defining a Problem**

For our project, it took a collaboration of five universities total, that presented progress records of many students from Freshman to Senior Year. For each student record, they have been marked “C” for completion of degree, “IP” for degree still in progress, and “D” to signify that student is a dropout. The goal overall would be to use Multiple Linear Regression to figure out the best indicators to predict a potential student of being on track to dropout.

* **Defining the Target Variable**

For this case study the target variable would be the “dropout” variable. Within the variable it will feature Y to represent the student has dropped out and N to represent that the student is within college still. We can convert this variable if needed into a binary variable as well to assist with our regression model when it comes time to creating the model.

* **Data Understanding**

Our data features many factors to determine a student’s status within their academic careers. This includes GPA, Year of Study, Program, and Number of Classes taken. Additionally, while reviewing the data we can see that everyone has a separate ID number to keep organization complete within the whole dataset. Lastly, the data is collected on a timeframe of 3 years, which is a very strong enough data for this project that carries across 5 Universities.

* **Data Preparation**

The dataset for this project did not require that much preparation. At first the team reviewed to see if there were any duplicate rows within the dataset, so it would not skew the data in a certain way. Additionally, there were sections in students records where GPAs were missing. Our team decided since there was less than 5% of the records that was missing the data, we used an imputation method of inserting the median of all GPA scores that were in the dataset. Another aspect of the dataset that needed to be changed was the athlete section where the student marked it Y or N. Our team adjusted the column to display 1s and 0s so it would help during the model creation process. Lastly, we removed a hobby column within the dataset because we felt as a team it would not provide any great insights or be a great predictor for the model we are creating.

* **Modeling**

For our modeling portion of this project, it required the building of our training and testing data sets. This is a requirement to create our prediction model itself. We decided to go with a 70:30 split of the data when creating the sets themselves. The two primary methods of prediction we were planning on using was multiple linear regression and K-Neighbor’s classifier methods. Additionally, if we wanted to look at a way to remove bias within the model as well, we could focus on a cross-validation split. Within the split we would use a stratified split, which assists with our bias issues if it were to occur when we run our model.

* **Model Interpretation**

For our model interpretation portion of the project, we are attempting to discover the factors or predictors that can helps understand where there may be a student that is a candidate to be the next dropout. This is done by comparing our trained models to see the accuracy within them to make a judgment call. Additionally, we can use an ROC curves to assist in the model selection process. Lastly, the batch training method would be the best approach for when creating our datasets. Once we have compared the different models our team has built, we can then go forward on to model deployment of the project once the best accuracy is confirmed from what we have studied.

* **Model Deployment**

After deployment has completed, American Public University System has assisted in decreasing the undergraduate dropout rate by a total of 5% percent. This would be a decrease from 40% to 35% dropout rate for undergrads. Additionally, the model can continue to build off the results and be ran after every semester or quarter depending on which Calendar format your universities use to develop more accurate results. Overall, our model was successful this go around to assist in keeping our youth within college; however, as with any model there is always room for improvement which our team will investigate going forward.

**References**

*College Dropout Rates*. (2021, June 25). EducationData. <https://educationdata.org/college-dropout-rates>

*Overview of Different Approaches to Deploying Machine Learning Models in Production*. (n.d.). KDnuggets. Retrieved July 5, 2021, from https://www.kdnuggets.com/2019/06/approaches-deploying-machine-learning-production.html

*School Dropout Rates in Iowa: State-wide and by Subgroup | mydata.iowa.gov*. (n.d.). Socrata. Retrieved July 5, 2021, from https://mydata.iowa.gov/Primary-Secondary-Ed/School-Dropout-Rates-in-Iowa-State-wide-and-by-Sub/negh-sf87/data