**Project Plan**

**Machine Learning**

**SEIS 763**

Dane Aadland

Jerome Duepner

Robert Krall

Sander Quint

Ken Tamura

**Data Source Description**

The data was obtained from the University of St. Thomas' student information system (Banner). The data contains a variety of information about GPS students, such as enrollment, demographic, and academic success. The data has been de-identified in order to maintain FERPA guidelines.

**Number of Records**

Our data contains 933 records.

**Number of Attributes**

Our data has 30 attributes.

**Attribute Descriptions**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Type** | **Definition** |
| Fake\_ID | Numerical | The data has been de-identified. This variable acts as a unique identifier for each record |
| STATUS\_DESC | Categorical | Tracks whether the student's degree has been awarded or is still in progress |
| PROGRAM\_DESC | Categorical | The overarching program in which the student is enrolled |
| MAJOR\_DESC | Categorical | The specific primary major in which the student is enrolled |
| SECOND\_MAJOR\_DESC | Categorical | The secondary major, if any, in which the student is enrolled |
| OUTCOME\_NUMBER | Numerical | Total number of degrees awarded to this specific student |
| OUTCOME\_AWARDED\_IND | Categorical | Denotes if the student was awarded the degree |
| CREDITS\_ATTEMPTED | Numerical | The number of credits the student attempted to earn while enrolled in this specific program |
| CREDITS\_EARNED | Numerical | The number of credits the student actually earned while in this specific program |
| GPA | Numerical | The grade point average (GPA) for the student while enrolled in this specific program |
| CITIZENSHIP\_DESC | Categorical | Citizenship status for the student |
| RELIGION\_DESC | Categorical | Self-reported religious preference for the student |
| VETERAN\_CATEGORY\_DESC | Categorical | Veteran status for the student |
| GENDER | Categorical | Self-reported gender for the student |
| PRIMARY\_ETHNICITY\_DESC | Categorical | Self-reported race/ethnicity for the student |
| MARITAL\_STATUS\_DESC | Categorical | Self-reported married status |
| CITY | Categorical | Student's city of residence |
| COUNTY\_DESC | Categorical | Student's county of residence |
| STATE\_PROVINCE | Categorical | Student's state (or province) of residence |
| POSTAL\_CODE | Categorical | Student's ZIP code of residence |
| NATION | Categorical | Student's nation of residence |
| PASSPORT\_ISSUE\_NATION\_DESC | Categorical | Nation that issued the immigration visa |
| IMMIGRATION\_STATUS | Categorical | Status of immigration visa |
| NATION\_OF\_BIRTH\_DESC | Categorical | Nation of birth |
| NATION\_OF\_CITIZENSHIP\_DESC | Categorical | Nation of citizenship |
| VISA\_TYPE\_DESC | Categorical | Type of immigration visa |
| GPS\_Start\_Term\_ | Categorical | Semester the student started taking courses in the GPS program |
| GPS\_Last\_Term\_ | Categorical | Last semester the student was registered for classes |
| Took\_Non\_GPS\_Course\_Within\_Year? | Categorical | Number of courses the student took outside of the GPS department |
| Retained? | Categorical (Y) | Indicator of whether the student was retained (registered for a course in the last year) |

|  |
| --- |
| **General Statistics** |
| * The dataset contains around six years' worth of information (from 2010-2016) |
| * Average GPA is 3.37 |
| * 54.3% of the students in the dataset are US citizens |
| * Students on average attempt 28.9 credits, but earn an average of 28.0 credits |
| * Students who graduated have earned an average of 34.3 credits |
| * Students who have yet to graduate have earned an average of 16.6 credits |
| * 303 students are female * 630 students are male |

**Tools**

|  |  |
| --- | --- |
| **Tool** | **Purpose** |
| Cognos | Used to query and obtain the dataset from the data store |
| Matlab | Used for data pre-processing, logistic regression, machine learning, and prediction |
| Python | Used for data pre-processing , logistic regression, machine learning, and prediction |
| R | Used for data pre-processing, logistic regression, machine learning, and prediction |
| Tableau | Used for data visualization |
| PowerPoint | Used for presenting the final project |

**Methods**

* Logistic Regression
* Support Vector Machine
* Gradient Boosting Machine
* Stacked Ensemble

**Problems & Questions**

* What are the types of students who most frequently leave their GPS program before graduating?
* What are the aspects of those students who complete their GPS degree?
* Can we predict and identify students who will be more likely not to complete their GPS degree?