**By Team 2**

Vaishali Siddeshwar

Niyanta Siddhapura

Rahul Kumar

Aritra Roy

Gagandeep Singh

Shivam Verma

CENTENNIAL COLLEGE

COMP 258-Neural Network PROJECT

Developing a predictive neural network model for HYPE dataset

Contents

[**1.** **Introduction** 1](#_Toc90443662)

[**1.1** **About the problem** 2](#_Toc90443663)

[**1.2** **The Dataset Source** 2](#_Toc90443664)

[**1.3** **General Information** 2](#_Toc90443665)

[**1.4** **Libraries used** 3](#_Toc90443666)

[**2.** **Data Analysis** 3](#_Toc90443667)

[**2.1.** **About Data** 3](#_Toc90443668)

[**2.2.** **Feature engineering** 4](#_Toc90443669)

[**2.3.** **Statistical Analysis** 4](#_Toc90443670)

[**2.4.** **Finding Answers to Business Questions** 6](#_Toc90443671)

[**2.4.1.** **What is the success rate among girls and boys?** 6](#_Toc90443672)

[**2.4.2.** **Which agegroup do most students belong to?** 7](#_Toc90443673)

[**2.4.3.** **Which agegroup has the highest Success and Unsuccess rate?** 7](#_Toc90443674)

[**2.4.4.** **Does Academic performance vary with Success rate?** 8](#_Toc90443675)

[**2.4.5.** **If Student did not persist during First Year Persistance, is the student more unsuccessful than successful?** 9](#_Toc90443676)

[**2.4.6.** **Are students belonging to a particular Applicant Category Name more Successful? Which Category is least successful?** 10](#_Toc90443677)

[**2.4.7.** **Does the Success Rate differ by Applicant Target Segment Name?** 11](#_Toc90443678)

[**1.** **Data Pipeline** 11](#_Toc90443679)

[**1.1.** **Feature Selection** 11](#_Toc90443680)

[**1.2.** **Data Pipeline** 12](#_Toc90443681)

[**2.** **Modelling** 13](#_Toc90443682)

[**2.1.** **Model Selection** 13](#_Toc90443683)

[**3.** **Fine-tuning** 14](#_Toc90443684)

[**4.** **Hosting** 14](#_Toc90443685)

[**5.** **Conclusion** 15](#_Toc90443686)

# **Introduction**

## **About the problem**

There is a need to develop a software product that can detect if a student succeeds in a course he/she has taken or not. If the likelihood of a student being unsuccessful in his/her course is detected early, then the college can provide him the needed support to help him succeed.

## **The Dataset Source**

The dataset was obtained from the course shell. This dataset contains student records.

## **General Information**

Original format: csv

Categorical Features

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **# Of Categories** |
| INTAKE COLLEGE EXPERIENCE | First time enrollment/Other enrollment status | 6 |
| SCHOOL CODE | Code | 7 |
| PROGRAM LONG NAME | Program name | 2 |
| STUDENT LEVEL NAME | Diploma or secondary | 2 |
| STUDENT TYPE NAME | New or returning | 1 |
| STUDENT TYPE GROUP NAME | Student group | 2 |
| TIME STATUS NAME | Fulltime or parttime student | 2 |
| RESIDENCY STATUS NAME | Resident or non-resident | 2 |
| FUNDING SOURCE NAME | Where did the funding come from | 5 |
| MAILING CITY NAME | City name | 29 |
| MAILING POSTAL CODE GROUP 3 | Postal code | 79 |
| MAILING POSTAL CODE | Postal code | 335 |
| MAILING PROVINCE NAME | Province name | 3 |
| GENDER | Gender | 2 |
| DISABILITY IND | Disability yes/no | 2 |
| MAILING COUNTRY NAME | Country | 1 |
| CURRENT STAY STATUS | Graduated/dropped out/studying | 10 |
| ACADEMIC PERFORMANCE | Grade | 4 |
| SUCCESS LEVEL | Successfully completed or not | 2 |
| FIRST YEAR PERSISTENCE COUNT | Persisted first year | 2 |
| HS AVERAGE GRADE | Grade | 9 |
| AGE GROUP LONG NAME | Bucketed Age group | 7 |
| FIRST GENERATION IND | First Generation or not | 2 |
| APPL FIRST LANGUAGE DESC | First language | 3 |
| APPL EDUC INST TYPE NAME | Education type | 1 |
| APPLICANT CATEGORY NAME | High school or mature student | 5 |
| APPLICANT TARGET SEGMENT NAME | Direct or international entry | 4 |
| PREV EDU CRED LEVEL NAME | High school or post-secondary | 3 |

**Numerical Features**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| ID 2 | Student ID |
| RECORD COUNT |  |
| PRIMARY PROGRAM CODE |  |
| PROGRAM SEMESTERS |  |
| TOTAL PROGRAM SEMESTERS |  |
| EXPECTED GRAD TERM CODE | Year/Month format |
| HS AVERAGE MARKS | Marks |
| ENGLISH TEST SCORE | Score |

**Composite Features**

FUTURE TERM ENROL: consists of 10 Boolean values of enrolled or not.

**Ordinal Numerical Features**

ADMIT TERM CODE:Year/Month

INTAKE TERM CODE: Year/Month

## **Libraries used**

sklearn

scipy

matplotlib

seaborn

Tensorflow

# **Data Analysis**

The code is in Group1\_SupervisedLearningProject/1.DataAnalysis.ipynb folder.

## **About Data**

This dataset contains student records. Following is the shape of the data



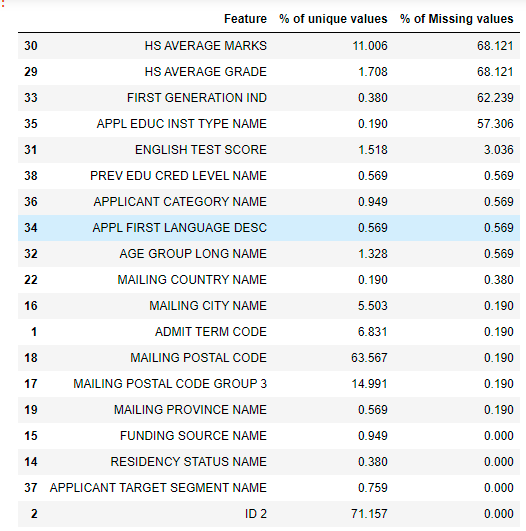
## **Feature engineering**

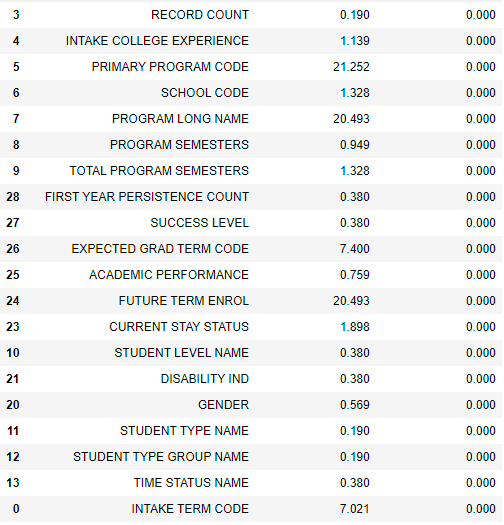
* Convert the year and month in ADMIT TERM CODE into Admit\_ODD\_orEVEN column. This column means if the student enrolled during an Odd or Even Term.
* We will also create a ADMIT YEAR column. This will help us to group the students based on ID and keep only the latest enrollment.
* we will convert the composite column \_FUTURE TERM ENROL into 10 different columns after splitting based on – character
* We found that CURRENT STAY STATUS has 10 categories - Left College - Enrolled 1 Term, Graduated - On Time, Left College - Enrolled 2 Terms, Graduated - Extended Time, Left College - Enrolled 3 Terms, Graduated - Different Program, Left College - Enrolled 5 or more Terms, Left College - Enrolled 4 Terms, Completed - Preparatory Program, Graduated - Different Intake.

We can reduce these categories to - Left College, Graduated, Completed

## **Statistical Analysis**

Next, let us compute the percentage of Missing values and the percentage of Unique values for all features in the data.





From the above output, following are our observations:

From the above output, following features will be dropped:

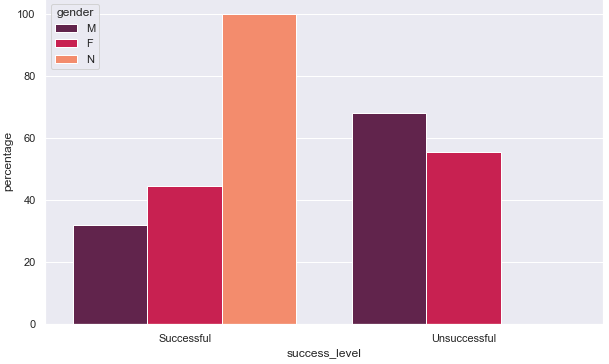
1. 'INTAKE TERM CODE', 'RECORD COUNT', 'PRIMARY PROGRAM CODE', 'STUDENT TYPE NAME', 'STUDENT TYPE GROUP NAME', 'MAILING CITY NAME', 'MAILING POSTAL CODE GROUP 3', 'MAILING POSTAL CODE', 'MAILING PROVINCE NAME', 'EXPECTED GRAD TERM CODE', 'FIRST GENERATION IND', 'APPL EDUC INST TYPE NAME', 'PROGRAM LONG NAME' will be dropped because they contain too many unique values or contain too many null values or they donot influence our predictor variable-SUCCESS LEVEL.
2. Drop APPL FIRST LANGUAGE DESC because 57% of values are UNKNOWN or Blank. Rest is English.
3. Impute MAILING COUNTRY NAME: We found that the rows that contain MAILING COUNTRY NAME as NA belong to CANADA. We found this using CITY NAME column.
4. Impute ENGLISH TEST SCORE: As this is a numeric field. We will impute using mean value.
5. Impute PREV EDU CRED LEVEL NAME: As this is a categorical field, we will impute this with mode.
6. Impute APPLICANT CATEGORY NAME: As this is a categorical field, we will impute this with mode.
7. Impute AGE GROUP LONG NAME: As this is a categorical field, we will impute this with mode.

We have analysed all the columns. We can use this logic to build the pipeline. The pipeline code is in 2.pipeline.ipynb.

## **Finding Answers to Business Questions**

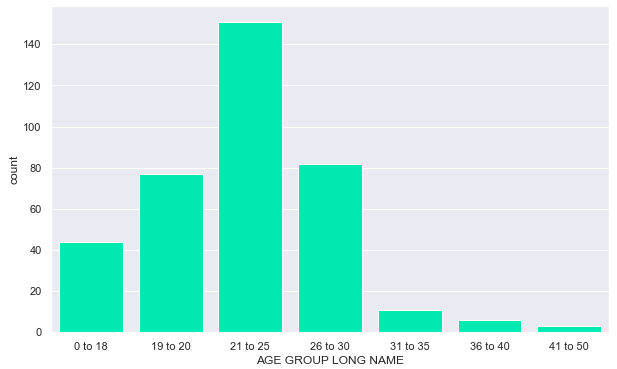
The code is in Group1\_SupervisedLearningProject/1.DataAnalysis.ipynb folder.

## **What is the success rate among girls and boys?**



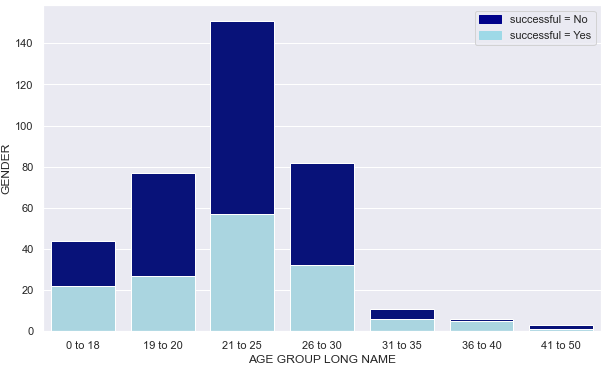
The above figure shows the success levels of Male and Females. There is only 1 record that contains N as gender, so let’s ignore N. Among those who are successful, Females counts(around 45% more) are more than males. On other hand, among those who are unsuccessful, Males are often more unsuccessful(more 65% more) than females.

## **Which agegroup do most students belong to?**



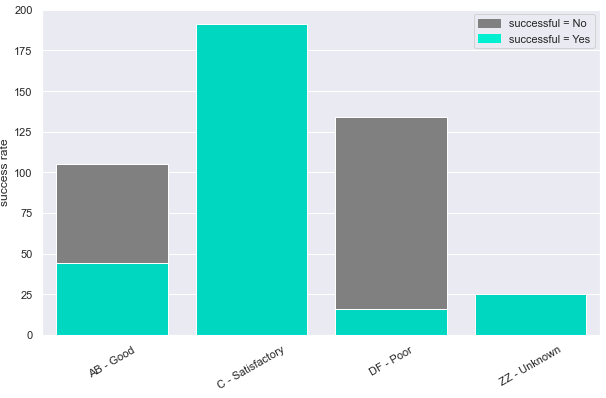
Most of the students belong to 21 to 15 age group followed by 26 to 30.

## **Which agegroup has the highest Success and Unsuccess rate?**



The above figure groups the data into different age groups and further groups each age-group by successful or unsuccessful. We can see that most of the students among 21-25 age group and 26-25 age group are most unsuccessful.

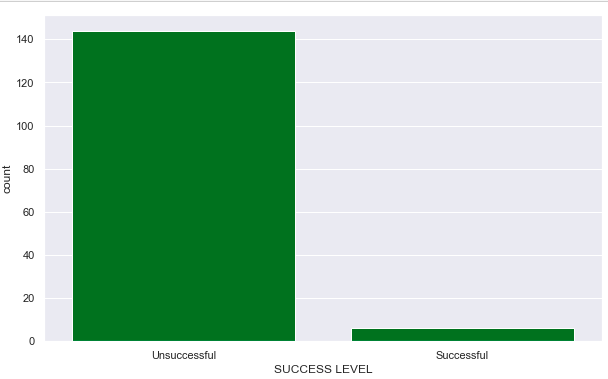
## **Does Academic performance vary with Success rate?**



From the above figure we can see that

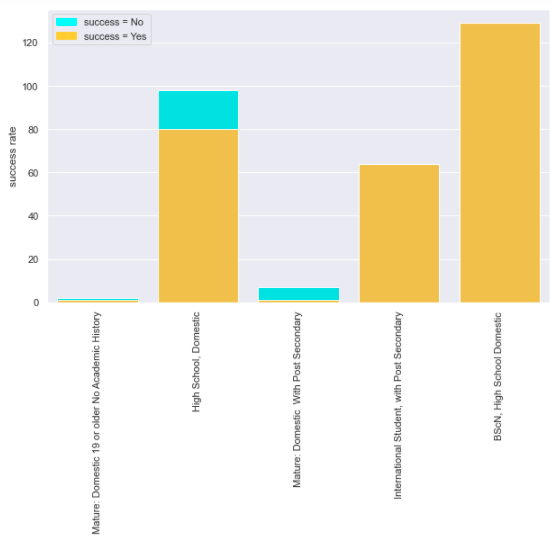
* Those who got A and B were half successful and half unsuccessful.
* Those who got C are mostly successful
* Those who got D and F are mostly unsuccessful.

## **If Student did not persist during First Year Persistance, is the student more unsuccessful than successful?**



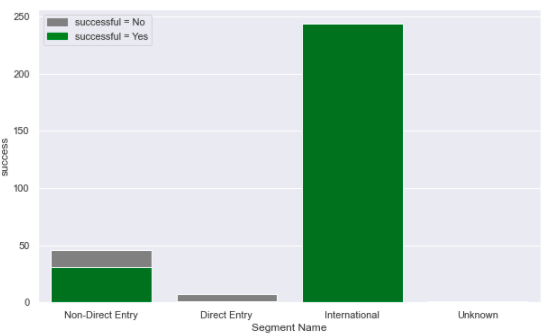
Yes, those who didn’t persist during the first year are most likely to be unsuccessful.

## **Are students belonging to a particular Applicant Category Name more Successful? Which Category is least successful?**



Those in International Post Secondary and BScN, High School Domestic are mostly successful. But, domestic high School students are noticeably unsuccessful.

## **Does the Success Rate differ by Applicant Target Segment Name?**



Let us ignore the Uknown Target. Most of the International Applicants are successful Some of Non-Direct Entry candidates are unsuccessful.

# **Data Pipeline**

## **Feature Selection**

The code is in Group2\_NeuralNetworksProject/featureSelection.pynb.

Following is the list of categorical columns that we need to filter through.

INTAKE COLLEGE EXPERIENCE

SCHOOL CODE

STUDENT LEVEL NAME

TIME STATUS NAME

RESIDENCY STATUS NAME

FUNDING SOURCE NAME

GENDER

DISABILITY IND

MAILING COUNTRY NAME

CURRENT STAY STATUS

ACADEMIC PERFORMANCE

FIRST YEAR PERSISTENCE COUNT

AGE GROUP LONG NAME

APPL FIRST LANGUAGE DESC

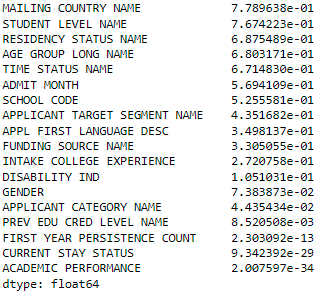
APPLICANT CATEGORY NAME

APPLICANT TARGET SEGMENT NAME

PREV EDU CRED LEVEL NAME

ADMIT MONTH

We have used Chi2Test to keep the top categorical features that appear to influence the Target. Following is the output of Chi2Test along with p-values



The above figure listed features in the in the decreasing order of p-values. A higher pvalue indicates a good dependence with the Target Variable. Seeing the pvalues, we can conclude that FIRST YEAR PERSISTENCE COUNT, CURRENT STAY STATUS and ACADEMIC PERFORMANCE features can be dropped because it has very low p-value. Below is the list of features that will be used to train the models.

## **Data Pipeline**

The code are in Group2\_NeuralNetworksProject/2.DataPipeline\_Type1.ipynb, Group2\_NeuralNetworksProject/2.DataPipeline\_Type2.ipynb and Group2\_NeuralNetworksProject/2.DataPipeline\_bestFeatures.ipynb notebooks.

Following numeric features will be used in the pipeline.

'PROGRAM SEMESTERS', 'TOTAL PROGRAM SEMESTERS', 'FIRST YEAR PERSISTENCE COUNT', 'ENGLISH TEST SCORE', 'Term1', 'Term2', 'Term3','Term4', 'Term5', 'Term6','Term7', 'Term8', 'Term9', 'Term10'

Following categorical features will be used in the pipeline.

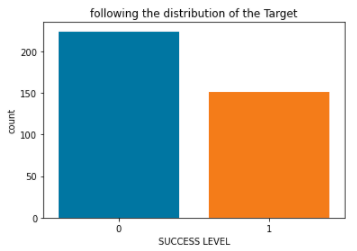
'INTAKE COLLEGE EXPERIENCE', 'SCHOOL CODE', 'STUDENT LEVEL NAME', 'TIME STATUS NAME',

'RESIDENCY STATUS NAME', 'FUNDING SOURCE NAME', 'GENDER', 'DISABILITY IND', 'MAILING COUNTRY NAME', 'CURRENT STAY STATUS', 'ACADEMIC PERFORMANCE', 'AGE GROUP LONG NAME', 'APPLICANT CATEGORY NAME', 'APPLICANT TARGET SEGMENT NAME', 'PREV EDU CRED LEVEL NAME'

The shapes are shown below



following is the distribution of the Target in the Train split. The ratio of Test Train split = 90%: 10%



We have a balanced data.

# **Modelling**

The code is in Group2\_NeuralNetworksProject/3.model\_allpandas\_type2.ipynb, 3.model\_allpandas\_typeq.ipynb and 3.model\_allpandas\_bestFeatures.ipynb files.

## **Model Selection**

Following is the list of NN models trained, their inputs, features used and their accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Input shape** | **Features used** | **Total params** | **Accuracy** |
| 1 Dense Layer of 32 Units, 1 DropOut 0.5 | (303, 70) | Removed 'APPL FIRST LANGUAGE DESC' | 2,305 | 0.85 |
| 1 Dense Layer of 32 Units, 1 DropOut 0.5 | (303, 76) | Included most of the features + Added Winter/Fall/Summer Sem feature | 2,497 | 0.95 |
| 1 Dense Layer of 32 Units, 1 DropOut 0.5 | (303, 68) | Dropped 3 categorical features with least p-value in Chi2Test  'FIRST YEAR PERSISTENCE COUNT’, 'CURRENT STAY STATUS’, ‘ACADEMIC PERFORMANCE', | 2,241 | 0.84 |

# **Fine-tuning**

In this section, we describe how we finetuned Random Forest classifier. The code is in Group1\_NeuralNetworkProject/4.Model\_finetune.ipynb folder.

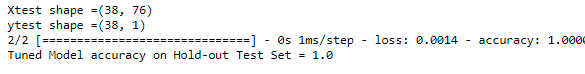
After tuning the model using Keras\_tuner for learning Rate and Number of Hidden Units in Dense Layer, we got the following output-



We used this to retrain our model and we got the following Accuracy on Train



Now, its time to evaluate the Best Model on the Unseen test Set-

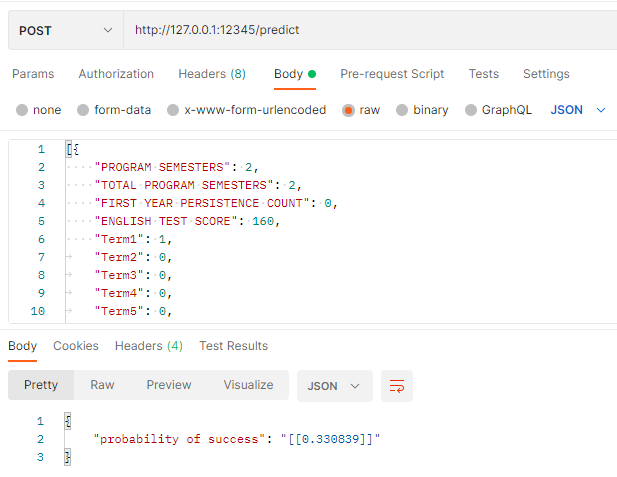


We had experimented with Random Forest Classifier on this data, we found this conventional Machine Learning model also has Accuracy of 1 on Unseen test Set.

# **Hosting**

The code is in Group2\_NeuralNetworkProject/5.hosting.py and utils.py files.

We hosted a flask application at URL -http://127.0.0.1:12345/predict that handles the POST requests. Below is a sample prediction output.



A sample request JSON is available in request.txt for reference.

# **Conclusion**

In this project we have demonstrated,

* exploratory data Analysis
* Select features using Chi2Test & build a pipeline to transform the data
* Model selection
* Fine-tune using Keras Tuner library
* Improvise the model’s Accuracy through by finetuning from 0.85 to 0.95.
* Hosting a neural network model on flask.