

# **Predictive Analytics Process Using Machine Learning for Student's Retention**

2021 TAIR Annual Conference (Virtual)

Luciano Boas

Rohan Patil

Abhishek Kumar

February 26th, 2021





# **01** Overview

- Introduction
- Machine Learning Basics

## 02 Exploratory Analysis

- Dataset
- Exploratory Analysis

## **03** Machine Learning

- Logistic Regression
- SVM
- KNN
- Comparison of Algorithms
- Conclusion and Future Directions

#### A PRELIMINARY STUDY ON RETENTION



"...eight years later, one-third of those individuals are no longer enrolled and have not earned any formal credentials." –National Student Clearinghouse.



#### PROJECT MANAGEMENT OVERVIEW



"First comes thought; then organization of that thought, into ideas and plans; then transformation of those plans into reality." - Napoleon Hill



#### MACHINE LEARNING (ML) APPROACHES/BEST PRACTICES



"Machine Learning learns patterns from historic data (training) and tries to generalize that in unseen (new) data." Mathangi Sri



# All Data Dataset Unseen data Train Validation Test Source: https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks

#### Common problems in the Training dataset (ML modeling):



#### **Best Practices (not limited to):**

- Check for missing data
- Check for collinearity
- Avoid data leakage
- Manage imbalanced classifier
- Don't reinvent the wheel
- Focus on model explainability
- Quick and dirty | Fail fast
- Minimum Viable Product (MVP)

# MACHINE LEARNING – TWO MOST COMMON TYPES OF LEARNING



Machine Learning allows testing how different algorithms perform when doing the same task.



#### RESOURCES



#### Machine Learning & Data Science Project Management

- <u>Structuring Machine Learning Projects</u> by DeepLearning.AI Coursera.
- The Machine Learning Canvas by Louis Dorard.
- The Practical Guide to Managing Data Science at Scale by Domino.
- <u>Team Data Science Process</u> by Microsoft.
- <u>TAMIDS Data Science Webinars</u> by Jian Tao, Texas A&M Institute of Data Science (TAMIDS).





# **01** Overview

- Introduction
- Machine Learning Basics

## **02** Exploratory Analysis

- Dataset
- Exploratory Analysis

## **03** Machine Learning

- Logistic Regression
- SVM
- KNN
- Comparison of Algorithms
- Conclusion and Future Directions

#### Dataset Attributes – Fall 2012 to Fall 2014



- No. of students retained = 9442
- No. of students not retained = 1919
- Total number of students = 11361
- Dependent Variable : RET1FF (Retention Fall to Fall)

Columns	Column Desc
ACAD_STAND	Academic Standing
AGE	Age
CITIZEN	Citizen
CRED_ATMP	Credits Attempted
EOT_GPA	End of Term GPA
ETHNICITY	Ethnicity
GENDER	Gender
GPA_CREDITS	GPA Credits
HOME	Home
HONORS	Honors
HS_GPA	High School GPA
HS_PRCTL	High School Percentile
HS_RANK	High School Rank
HS_SIZE	High School Size
PART_FULL	Part Full
QL_PTS	Quality Points
RET1_FS	Retention Fall to Spring
Sat_super_Score	SAT Super Score
STUD CLASS	Student Classification

# We Coded categorical data into numerical values

Columns	Coded Values
Part Full	$\mathbf{F}=1 \ , \ \mathbf{P}=0$
Student Classification	FR = 1, Others = 0
Ethnicity	B = 0, $WH = 1$ , $HI = 2$ , $A = 3$ , $All$ Others = 4
Gender	M = 1 , $F = 0$
Home	Instate = 1, Others 0
Academic Standing	GS = 1, $SP = 2$ , Else 0
Honors	Y = 1 Else 0
AGE	LTE 17 = 0, EQ 18 = 1, GTE 19 = 2

\*Null Values were removed from the dataset

#### **Dataset Distribution**



#### "Histogram is a plot that lets you show the underlying frequency distribution or the **probability distribution** of a **continuous numerical variable**."



# 

- 1.0

- 0.8

#### **Correlation Matrix**

ACAD_STAND	- 1	-0.0098	-0.0018	0.093	0.75	-0.045	0.16	0.13	-0.028	0.1	0.26	0.28	-0.14	0.021	0.064	0.69	0.16	0.085	-0.028	0.32
AGE	0.0098	1	0.034	-0.048	-0.028	-0.07	-0.13	-0.043	0.026	-0.024	-0.02	-0.045	-0.0014	-0.035	-0.0049	-0.043	0.0027	-0.05	0.018	0.0032
CITIZEN	0.0018	0.034	1	-0.04	-0.026	0.3	0.067	-0.042	0.26	0.01	0.066	0.0036	0.013	0.046	0.006	-0.037	-0.018	0.033	-0.017	-0.019
CRED_ATMP	- 0.093	-0.048	-0.04	1	0.21	-0.056	0.024	0.65	-0.06	0.12	0.12	0.16	-0.096	-0.016	0.3	0.42	0.079	0.22	0.02	0.059
EOT_GPA	0.75	-0.028	-0.026	0.21	1	-0.077	0.21	0.34	-0.043	0.2	0.36	0.39	-0.2	0.026	0.087	0.92	0.27	0.2	-0.039	0.37
ETHNICITY	0.045	-0.07	0.3	-0.056	-0.077	1	0.013	-0.011	0.095	-0.08	-0.06	-0.0052	-0.01	0.021	0.015	-0.072	-0.0048	-0.19	-0.017	-0.02
GENDER	0.16	-0.13	0.067	0.024	0.21	0.013	1	0.061	0.024	0.053	0.17	0.18	-0.087	0.018	0.021	0.2	0.029	-0.12	-0.038	0.066
GPA_CREDITS	0.13	-0.043	-0.042	0.65	0.34	-0.011	0.061	1	-0.046	0.11	0.14	0.17	-0.089	0.0069	0.22	0.6	0.2	0.14	-0.00093	0.16
HOME	0.028	0.026	0.26	-0.06	-0.043	0.095	0.024	-0.046	1	-0.035	0.099	-0.029	0.13	0.17	0.0049	-0.056	0.0053	-0.053	-0.008	0.0072
HONORS	- 0.1	-0.024	0.01	0.12	0.2	-0.08	0.053	0.11	-0.035	1	0.21	0.25	-0.17	-0.037	0.015	0.22	0.064	0.38	-0.037	0.098
HS_GPA	0.26	-0.02	0.066	0.12	0.36	-0.06	0.17	0.14	0.099	0.21	1	0.64	-0.35	-0.06	0.04	0.36	0.05	0.2	-0.091	0.13
HS_PRCTL	0.28	-0.045	0.0036	0.16	0.39	-0.0052	0.18	0.17	-0.029	0.25	0.64	1	-0.65	-0.15	0.052	0.4	0.065	0.23	-0.12	0.14
HS_RANK	0.14	-0.0014	0.013	-0.096	-0.2	-0.01	-0.087	-0.089	0.13	-0.17	-0.35	-0.65	1	0.74	-0.015	-0.21	-0.022	-0.11	0.16	-0.058
HS_SIZE	- 0.021	-0.035	0.046	-0.016	0.026	0.021	0.018	0.0069	0.17	-0.037	-0.06	-0.15	0.74	1	0.019	0.02	0.024	0.034	0.17	0.032
PART_FULL	- 0.064	-0.0049	0.006	0.3	0.087	0.015	0.021	0.22	0.0049	0.015	0.04	0.052	-0.015	0.019	1	0.14	0.066	0.05	-0.014	0.064
QL_PTS	0.69	-0.043	-0.037	0.42	0.92	-0.072	0.2	0.6	-0.056	0.22	0.36	0.4	-0.21	0.02	0.14	1	0.26	0.24	-0.032	0.35
RET1_FS	0.16	0.0027	-0.018	0.079	0.27	-0.0048	0.029	0.2	0.0053	0.064	0.05	0.065	-0.022	0.024	0.066	0.26	1	0.042	-0.0068	0.49
Sat_super_Score	- 0.085	-0.05	0.033	0.22	0.2	-0.19	-0.12	0.14	-0.053	0.38	0.2	0.23	-0.11	0.034	0.05	0.24	0.042	1	-0.0086	0.04
STUD_CLASS	0.028	0.018	-0.017	0.02	-0.039	-0.017	-0.038	-0.00093	-0.008	-0.037	-0.091	-0.12	0.16	0.17	-0.014	-0.032	-0.0068	-0.0086	1	-0.014
RET1_FF	0.32	0.0032	-0.019	0.059	0.37	-0.02	0.066	0.16	0.0072	0.098	0.13	0.14	-0.058	0.032	0.064	0.35	0.49	0.04	-0.014	1
	ACAD_STAND .	AGE	CITIZEN .	CRED_ATMP	EOT_GPA	ETHNICITY .	GENDER .	GPA_CREDITS	HOME	HONORS .	HS_GPA	HS_PRCTL	HS_RANK	HS_SIZE	PART_FULL	OL_PTS .	RET1_FS -	Sat_super_Score .	STUD_CLASS	RET1_FF

- 0.6 - 0.4 - 0.2 - 0.0

- -0.6

- -0.4



Top reasons to use feature selection are:

"It enables the machine learning algorithm to train faster."

"It reduces the complexity of a model and makes it easier to interpret."

"It improves the accuracy of a model if the right subset is chosen."

"It reduces overfitting."



Source: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/#:~:text=Top%20reasons%20to%20use%20feature, the%20right%20subset%20is%20chosen.

#### Important Features with Ret1FF



"Pairplot shows the relationship for (n,2) combination of variable in a DataFrame as a matrix of plots and the diagonal plots are the univariate plots."



https://www.tutorialspoint.com/seaborn/seaborn\_visualizing\_pairwise\_relationship.htm

#### **Confusion** Matrix

#### True Positive:

"Interpretation: You predicted positive and it's true. You predicted that the student will be retained, and is retained"

#### **True Negative:**

"Interpretation: You predicted negative and it's true. You predicted that the student will not be retained and is not retained"

#### False Positive: (Type 1 Error)

"Interpretation: You predicted positive and it's false. You predicted that the student will be retained when is not retained

#### False Negative: (Type 2 Error)

"Interpretation: You predicted negative and it's false. You predicted that the student will not be retained when is retained.

#### Actual Values







Dependent Variable	Precision	Recall	F1 - Score	Support
0 (Not Retained)	TN / (TN + FN)	TN / (TN + FP)	AVG(Precision, Recall)	No of Obs
1 (Retained)	TP / (TP + FP)	TP/(TP + FN)	AVG(Precision, Recall)	No of Obs

#### **Precision:**

"The precision returns the proportion of true positives among all the values predicted as positive."

#### **Recall:**

"The recall returns the proportion of positive values correctly predicted."



#### F1-score:

"The f1-score is the harmonic mean of precision and recall. It is often used to compare classifiers."

Support: "number of observations for each class."





## **01** Overview

- Introduction
- Machine Learning Basics

## **D2** Exploratory Analysis

- Dataset
- Exploratory Analysis

## **03** Machine Learning

- Logistic Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbor (KNN)
- Comparison of Algorithms
- Conclusion and Future Directions

#### LOGISTIC REGRESSION



- Linear regression: used to predict outputs on a continuous spectrum
- Logistic regression: used to predict binary outputs with two possible values labeled "0" or "1"
- Logistic model can be one of the two classes: pass/fail, win/loss, retained/not retained



Hours Studying	Pass/Fail
1	0
1.5	0
2	0
3	1
3.25	0
4	1
5	1
6	1

• <u>Linear equation:</u> •  $y = b_0 + b_1 * x$ 

•

Apply Sigmoid function: • P(x) = sigmoid(y)•  $P(x) = \frac{1}{1+e^{-y}}$ •  $P(x) = \frac{1}{1+e^{-(b_0+b_1*x)}}$ 

#### THE CLASS IMBALANCE PROBLEM



- A common problem in Machine Learning
- Almost all the instances belong to one major class and the rest to minor class



#### **Create balance though sampling**

#### **Re-sampling the data:**

- Over-sampling increases the number of minority class members
- Under-sampling aims to reduce the number of majority samples
- We used SMOTE for our preliminary research





	Precision	Recall	F1-score	Support
Not Retained	-	-	0.37	-
Retained	-	-	0.92	-

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.50	-
Retained	-	-	0.93	-

Without sampling and without feature selection

Without sampling and with feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.52	-
Retained	-	-	0.93	-

With oversampling and without feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.53	•
Retained	-	-	0.88	-

With oversampling and with feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.41	-
Retained	-	-	0.77	-

With undersampling and without feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.44	-
Retained	-	-	0.77	-

With undersampling and with feature selection

#### SUPPORT VECTOR MACHINES (SVM): MODEL EVALUATION



- SVMs are based on the idea of finding a hyperplane that divides a dataset into two classes
- The line that maximizes the minimum margin is a good bet.

FEATURE #2

• This maximum-margin separator is determined by a subset of the datapoints.



#### K NEAREST NEIGHBORS (KNN): MODEL EVALUATION



- k-nearest neighbors algorithm (KNN) is a classification algorithm
- KNN works by finding the **most similar** data points in the training data, and attempt to make an **educated guess** based on their classifications



R. Ahmed, M. Bouchard ML Classification Bootcamp in Python

#### SVM and KNN – MODEL TRAINING & TESTING



- train\_test\_split is a function in Sklearn model selection for splitting data arrays into two subsets
- train\_test\_split: 70:30

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.39	-
Retained	-	-	0.92	-

SVM - Without sampling and without feature selection

	Precision	Recall	F1-score	Support
Not Retained	-	-	0.44	-
Retained	-	-	0.89	-

KNN - Without sampling and without feature selection



Each model will have different performance characteristics

- **Prepare Dataset**. Load the libraries and dataset ready to train the models
- Train Models. Train standard machine learning models on the dataset ready for evaluation
- **Compare Models**. Compare the trained models using k-fold cross validation procedure



ML Models	Results
Logistic Reasoning	87.93% (0.079328)
Support Vector Machine	82.98% (0.112209)
K Nearest Neighbor	81.05% (0.104777)



	Precision	Recall	F1-score	Support	
Not Retained	-	-	0.45	•	
Retained	-	-	0.95		

**Classification Report for Fall 2015** 

#### **Conclusion and Future Directions**

- Logistic regression with oversampling of our data and feature selection produced best results
- Update our preliminary research with more salient parameters
- Use of other techniques and methods to handle imbalance datasets
- Use different machine learning models to predict various possibilities with HigherEd data



#### TEXAS TECH UNIVERSITY<sup>°</sup> From here, it's possible.