Prediction of university student attrition rate using Ridge and Lasso Regression

by

Teja Usha Sree Vallabhaneni

B.Tech., Koneru Lakshmaiah University, India, 2016

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Dr. William H. Hsu

Copyright

© Teja Usha Sree Vallabhaneni 2019.

Abstract

One of the major challenges faced by many institutions is the attrition rate. *Institutional attrition* is the phenomenon of individuals moving out of an institution, prior to completing term-limited programs; this term can apply to employees (e.g., postdoctoral fellows) or students (Bani, J., Haji, & M., https://pdfs.semanticscholar.org/94b1/, 2017). In the context of this project, which focuses on student attrition, it includes students who drop out, are dismissed, or do not return to their studies before the completion of their degree. The student attrition rate at a university is often measured in terms of net change in enrollment per year due to students discontinuing their studies at that university. One of the consequences of attrition is that students are unable to graduate despite significant investments in the form of funding from scholarship-granting institutions or governments.

This project deals with the study of factors contributing to student attrition rate at a land-grant state university and predicting whether a student is going to drop out or not based on various factors such as gender, race, cumulative GPA, etc. One reason that this study is timely and necessary is that a predictive model may allow an institution to recognize factors contributing to dropping out and will help the institution retain students and prevent dropout and "stopout" in certain cases. A decrease in preventable attrition may similarly enable more students to earn the degrees they were pursuing at a point where they can realize more of the professional and economic benefits of that degree.

The report comprises a brief review of the supporting literature for the task of student attrition rate prediction and describes a machine learning and data science project centered around further explorations of a previously-developed experimental test bed. These involve extraction of data from historical archives (raw data from the university registrar's office and other sources), cleaning the data, building the testing and training data sets for the supervised learning algorithms, training, and evaluation of models, and review of the models to derive actionable insights. Logistic regression, a supervised inductive learning algorithm, is used to train a classification model, which in turn is used to predict student dropout on a case wise basis. Regression models that use L2 regularization (ridge regression) and L1 regularization (lasso regression) will also be used to predict student dropout. These algorithms are used in feature selection and in the creation of a flexible model when data consists of a large set of features. Performance metrics such as the precision, accuracy, recall, and F1 score are used to compare the performance.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my committee members Dr. William H. Hsu, Dr. Mitchell Nielsen, and Dr. Doina Caragea for taking the time to serve on my committee and their support in the process of this project.

I am especially indebted to my major advisor Dr. William H. Hsu for believing in my abilities and for his constant support from my very first semester here at Kansas State University.

I am also grateful to Dr. Rodney Howell, Dr. Dennis Lang, and Dr. Jorge Valenzuela for being amazing mentors during my semesters as a teaching assistant.

I would like to extend a huge thanks to all my family members without whom this journey would not have been possible. I am extremely grateful to my mom, Rama Kumari Vallabhaneni, for her unconditional love, encouragement, and emotional support in my life. I cannot thank my sister, Sindhuri Vallabhaneni enough for her immense love and support. She is my strength and has played a significant role in what I have achieved so far in my life.

And last, but in no way least, I would like to thank my friends Sindhu Velumula, Sudesh Kumar Venkatramolla, Japneet Kaur Brar, Yojitha Chinthareddy, Pavan Manepalli, Poojitha Bikki, Pruthvidhar Dhodda, Sharmila Vegesana, and Sneha Gullapalli for their support and friendship. The past two years would not have been so memorable and fun without all of the////

Chapter 1 - Introduction

This chapter presents a brief overview of this project, starting with the problem statement, overall goal, and objectives and a synopsis of this project. It gives an overview of relevant extant methods for the data science application task of predicting institutional attrition and a rationale for the experimental approach taken and the evaluation methods used.

1.1 Problem Definition

Institutional attrition is the phenomenon of individuals moving out of an institution, prior to completing term-limited programs; this term can apply to employees (e.g., postdoctoral fellows) or students. In the context of this project, which focuses on student attrition, it means the number of students dropping out before completion of their degree. The student attrition rate at a university is often measured in terms of net change in enrollment per year due to students discontinuing their studies at that university. According to a study, one-third of the students who are expecting to graduate with a degree in higher education are dropping out due to personal and professional scenarios. 51% of the undergraduate students in the United States are leaving the school without a degree.

The increased attrition rate of the institutions is affecting the standard of living of people and investments of the various organization to achieve a good literacy rate. Students who enroll in a degree devote a lot of effort and investment to achieve the degree (Bonham, A, Luckie, I., & A., 1993; Bonham, A, Luckie, I., & A., 1993). When the students drop out before graduating or receiving a degree it counts as wastage of investment, as a missed opportunity to have a better standard of living and independent life (Johnson, 2012).

When we consider the public and the institutional investments, they are trying to make education accessible by providing services in the form of financial aids and improving the resources for education. When the student leaves the institution without a degree the goal of these investments is never achieved (DesJardins, L., McCall, & P., 2010)

Positive effects of decreasing student attrition rates include improvement of the institution's reputation and income. It also helps in decreasing the wastage of funds that are being invested in admitting students who are never receiving a degree. Thus, the prediction of institutional attrition rate helps the institutions in retaining the students and developing the strategies to help the students in achieving a degree. However, the prediction task is not always a hundred percent accurate since a student dropping out may involve several reasons like social, economic, psychological and sudden scenarios that may not arise in every other student case. But when the student data of at least several terms is studied, we may be able to notice the trends in various common scenarios that are feeding the student attrition rate. By taking measures to control these main factors contributing to student attrition rate the student retention can be increased considerably. In this project, the institutional attrition rate is predicted using various machine learning techniques with a focus on feature selection.

1.2 Goals and Technical Objectives

The goal of this project is to build a learning model that can predict the attrition rate of an institution using various classification techniques. The project also aimed at improvising the results by applying the regularization techniques on the data set. The data used for this project is provided by the Kansas State University and consists of the student academic data including few personal factors that are required to carry out education in an institution. The data is provided with an additional feature called drop out and the drop out is classified into “yes” or “No” classes.

The attrition was predicted based on many factors provided in the data set like the cumulative GPA, the permanent location of the student, the location of the campus, the housing of the student, etc. The classification techniques like Logistic regression and Regression techniques like Linear regression were applied to the data set to predict the attrition level. The regularization techniques like the Lasso and Ridge regression are used to solve the problems of overfitting and Feature selection. The feature selection is one of the major goals of this project. Feature selection is performed using various techniques to improvise the results.

1.3 Synopsis

The report comprises a brief review of the supporting literature for the task of student attrition rate prediction and describes a machine learning and data science project centered around further explorations of a previously-developed experimental test bed. These involve extraction of data from historical archives (raw data from the university registrar's office and other sources), cleaning the data, building the testing and training data sets for the supervised learning algorithms, training, and evaluation of models, and review of the models to derive actionable insights. Various features in the data set like Cumulative GPA, permanent location (in-state or from outside Kansas), Housing (on-campus or off campus),Number of credits opted for a semester, campus location is considered, and feature selection was carried out to select the features that play a major role in predicting the attrition. The data is partitioned into testing and training data sets. Logistic regression, a supervised inductive learning algorithm, is used to train a classification model, which in turn is used to predict student dropout on a case wise basis. Regression models that use the L2 and L1 regularization techniques (part of Ridge and Lasso Regression, respectively) will also be used to predict student dropout. These algorithms are used in feature selection and in the creation of a flexible model when data consists of a large set of features.

Chapter 2 - Background and Related Work

This chapter discusses a brief overview of the machine learning techniques used in this project, starting with the Literature survey on Classification Problem followed by Established methods of solving the Classification problems like Logistic Regression, Linear Regression, Lasso Regression, and Ridge Regression. The students who leave the institution prior to completion of the degree are classified into stop out and drop out students where drop out is the criteria considered in this project. Drop out indicates the students who did not receive a degree, students who did not enroll for the coming 6 semesters and not dismissed by the university or students whose data is not present in the end of the semester files. However, there is another scenario described as stop out which is described as students who take a temporary break which may be longer than expected but they plan to return and complete the degree in future. Lack of money or time are two main factors contributing in every 2 students out of 5 stop outs (Bonham & Luckie, 1993).The Drop out behavior is analyzed and predicted in the following project and Stop out behavior is considered for the future work.

2.1 Literature Survey: Classification Problem

Supervised learning is a machine learning algorithm technique which uses the labeled data to learn the mapping function between the input variable (x) and output variable (y) (Pedregosa, n.d.). The aim of this algorithm is to predict the output variable (y) when a new input data (x) which is not used in training is given. The algorithm using the trained data learns the mapping function to predict the output. The mapping function equation looks like the following equation. Where y is the output to be predicted, x is the input variable, f is the mapping function.

$$y = f(x)$$

Equation 2.1 Mapping Function

The supervised learning is mainly divided into classification and Regression tasks. Classification is one of the machine learning techniques that is used to predict the output variable which is discrete valued. An example is to identify whether an email is a spam or not. Regression is the other machine learning technique that is used in the prediction of output which is continuous valued. An example is to find the age of a person. Algorithms are trained on the Training data set and the resulting model is applied on the input from testing data set to predict the output. The classification model is trained on the well-labeled data and the model is further used to predict to which class the output belongs whether it is "1" (drop out) or "0" (not dropped out).

2.2 Established Classification Methods

2.2.1 Logistic Regression

The logistic regression is a regression analysis that uses the logistic function to map the one or more nominal, ordinal or interval independent variable to a dependent categorical variable. The dependent output variable has categories like "yes" or "No", "1" or "0", "drop out" or "not dropped out". The input variables can be independent and the values for the input variables can be of any type. Logistic regression generally falls into 3 categories. They are:

1. Binary Logistic Regression
2. Multinomial Logistic Regression
3. Ordinal Logistic Regression

Binary Logistic Regression: In this, the output variable should have dichotomous values like "yes" or "No".

Ex: To Identify whether an email is "Spam" or "Not Spam".

Multinomial Logistic Regression: In this, the output variables can fall into 3 or more categories like "Kid"," Young", "old".

Ex: To Identify whether a person is "vegan" or "Non-vegetarian" or "Vegetarian".

Ordinal Logistic Regression: The output variables generally fall into 3 or more categories that are ordered.

Ex: To Identify groups ranging between 1 to 5

The logistic regression generates the coefficients in the formula to estimate the possibility of the presence of this feature in the estimation of the output variable

$$\boldsymbol{Logit(p) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n}$$

Equation 2.2 Logistic Regression function

Where p is the probability that the output variable is dependent on the feature. $a_0, a_1, a_2, \ldots$ Are the coefficients for each feature in the training data. The logit function is the logged odds.

$$\boldsymbol{Odds = p/(1-p)}$$

Equation 2.3 Estimating Odds

The odds are estimated as the ratio of the probability of the presence of characteristic to the probability of absence of the characteristic.

$$\boldsymbol{Logit(p) = ln(p/(1-p))}$$

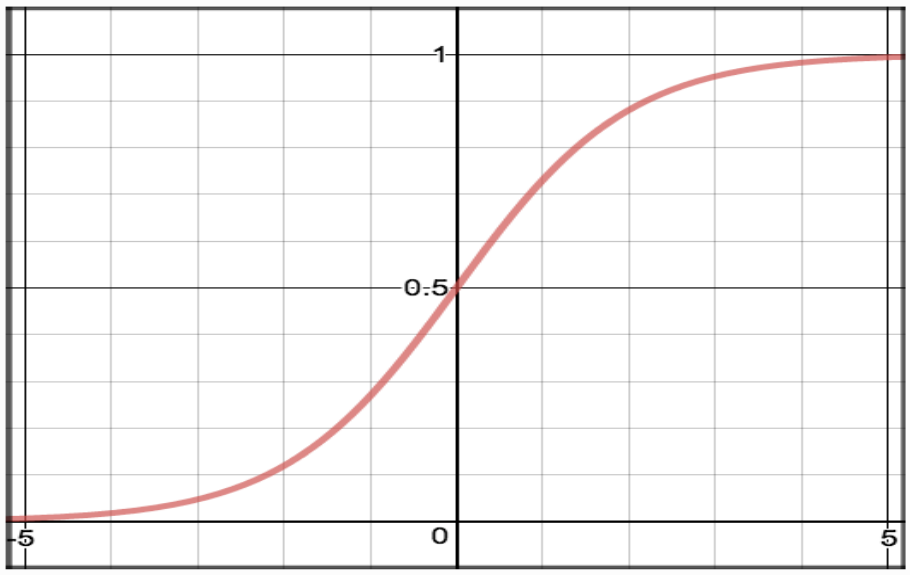
Equation 2.4 Logit Function for Logged Odds

In machine learning, a sigmoid function is used to map the predictions to the probabilities. This function helps in mapping any real values to real values in the range 0 to 1.

$$S(z) = 1/1 + e - Z$$

Equation 2.5 Sigmoid Function

S(z) is the output variable that needs to be predicted using the algorithm. The z is the input variables for the function. E is the base of the natural log.



(Aubele, n.d.)

Figure 2.1 Logistic Regression

The current mapping function returns the probability in the range of 0 to 1. We choose a tripping point and classify the observations above this point into class 1 and others into class 2.

2.2.2. Linear Regression

Linear Regression is one of the popular methods for supervised machine learning used to establish a relationship between continuous variables. This relation is mostly the statistical relation but not deterministic relation. In the deterministic relation, the relation between the dependent and the independent variable can be exactly established, unlike the statistical relation. The linear

function used to map the dependent variable from the input variables can be estimated as following where y is the dependent variable and x is the independent variable. a_0, a_1 are the coefficient and e are assumed to be the error.

$$\boldsymbol{y = a_0 + a_1 x + e}$$

Equation 2.6 Linear Function

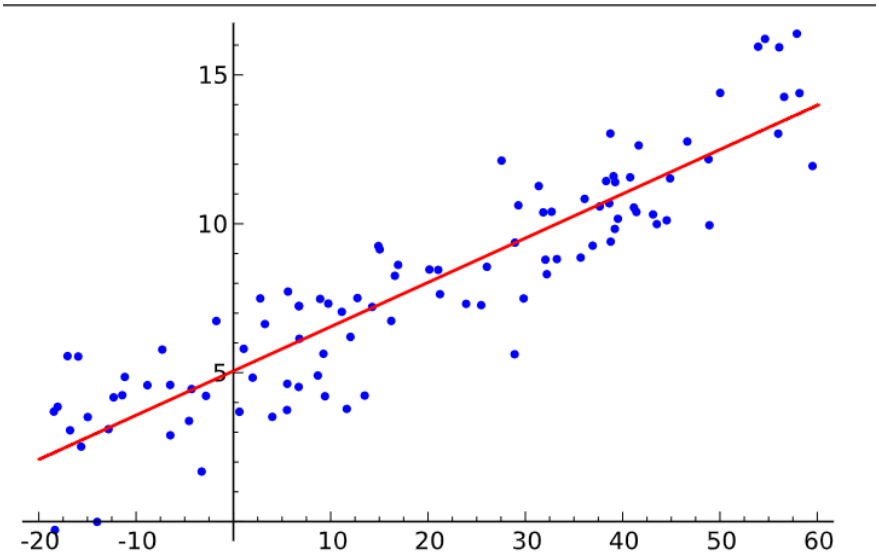
This is the simple Linear Regression equation. The Linear Regression is classified into 2 types. They are:

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression: In this type of regression one independent variable is predicted based on a single dependent variable.

Multiple Linear Regression: In this type of regression one dependent variable is predicted based on multiple independent variables.

The main aim of the Linear Regression is to find the best estimates for the coefficients such that the error between the predicted variable value to the actual variable value should be reduced. Instead of trails, directly linear algebra is used to predict the dependent variable. The linear Regression assumes that the independent variables are not correlated. R square is the metric that explains the percentage of variance explained by covariates in a model that has ranged between 0 to 1. Error metrics like RMSE, MSE, MAE are used in evaluation which has to be low.



(https://en.wikipedia.org/wiki/Linear_regression, 2019)

Figure 2.2 Linear Regression

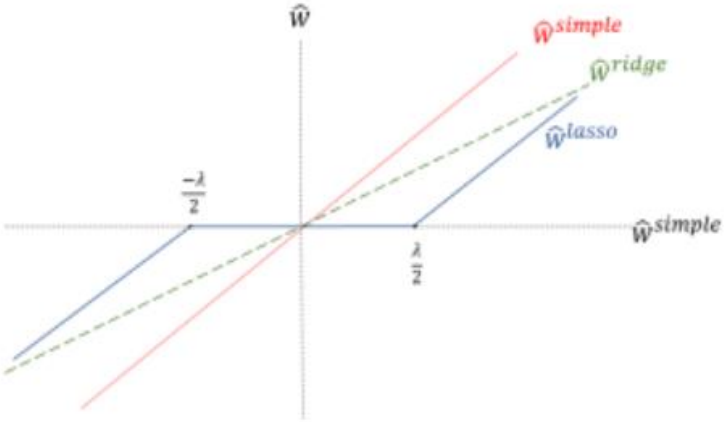
2.2.3 Lasso Regression

Lasso regression is explained as Least absolute shrinkage and Selection operator (Saptashwa, n.d.). Lasso regression is a method of regression analysis that helps in Feature selection and a regularization technique. This model is used to improve the prediction accuracy using the L1 regularization technique. Regularization techniques like Lasso regression are used to solve the problem of overfitting. Overfitting happens when the model learns from the noise and signal in the training data and becomes a bad predicting model for the testing data i.e. The data on which it is not trained. Regularization is used when there is complexity in the model and the parameters should be penalized. Regularization penalizes the parameters so that they won't overfit. It adds the "absolute values of magnitude" of coefficient as penalty term to the loss function.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=0}^{p}|\beta_j|$$

Equation 2.7 Lasso Regression

The main thing that lasso regression does is decreases the lambda value for the less important features and sometimes make the lambda value to 0 if the features are not playing any role in predicting the dependent variable.



(Jain S., 2019)

Figure 2.3 Plotting Lasso Regression coefficient values

The coefficients become 0 for a certain range and are reduced by a certain ratio which is why the lasso regression has relatively lower coefficients range compared to ridge regression.

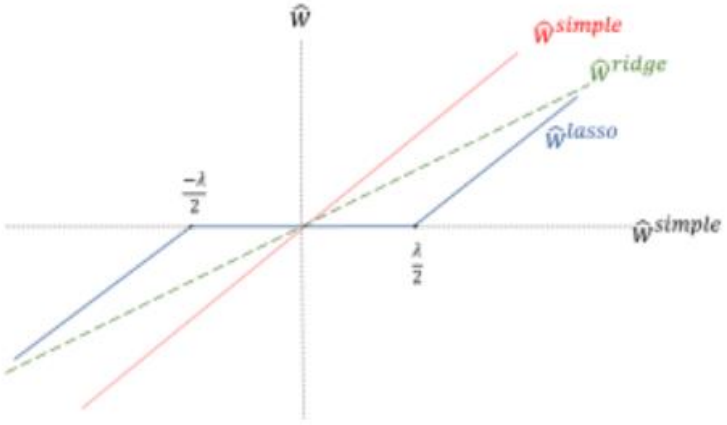
2.2.4 Ridge Regression

Ridge Regression is also another regularization technique that uses the L2 regularization technique. Ridge regression adds the “squared magnitude of the coefficient” as a penalty to the loss function. If the lambda value is too large it adds importance to the feature resulting in underfitting solves the problem of overfitting.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=0}^{p}{\beta_j}^2$$

Equation 2.8 Equation of Ridge Regression

The features importance can be reduced by removing the features completely by forwarding or backward selection or setting their value to 0. By this way, we can never predict the effect of the removed variable on the dependent variable. So, Ridge regression helps to set the value of the coefficient of these features if they are far from the common values. Since it shrinks the parameters it is used to prevent the multicollinearity.



(Jain, 2019)

Figure 2.4 Plotting of Ridge Regression coefficient values

Chapter 3 - Implementation

This chapter includes an overview of the data, data preprocessing methods and implementation steps of this work.

3.1 Overview of the Data

The training of a machine learning model highly depends on the training data supplied for the algorithm. In most cases, if the model performs misclassification or is not able to reach the expected accuracy then the reason will be noise in the data. The data should be preprocessed and should be made free from the noise and the null values. The null values should be handled by the appropriate techniques. In most of the cases, the rows containing the null values can be omitted if they are considered in the low count, null values can also be replaced by the mean/median/mode, the prediction of the null values can be carried out or the algorithms which support missing values can be considered.

The data set used for this project is obtained from Kansas state university. The data includes 3 different sets from semesters fall 2012 – spring 2018. The data includes the beginning of the semester records, end of the semester data and the graduated student data. Mostly the missing values were able to be replaced with values from the data at the end of the semester.

The student data set includes 15 different variables about the student details like a number of credits opted in each semester, permanent location, type of housing, graduation details, campus details, and the cumulative GPA, etc. The additional column called a drop out that needs to be predicted is added to the data set. The data set has 215,183 rows of records.

Each Student in the student data set consists of the following features.

- **Term:** The term Code in which the student is enrolled or the present enrolled term Code i.e. 2125 indicates semester 1. 2132 is semester 2
- **Descr:** The description of term i.e. Fall/Spring
- **Campus ID:** The unique identification number for the student
- **Perm State:** The state of the permanent location.
- **Housing:** This term indicates whether the student resides on campus or off campus
- **UGRD Admit Term:** The under graduation admit term
- **GRAD Admit Term**: The graduation admit term for the student
- **Readmit**: If the student readmitted into the university after a break.
- **Academic Plan**: type of degree or the academic program the student is enrolled into
- **CUM GPA:** The cumulative GPA of the student up to the date.
- **Semester Hours:** This explains the number of credit hours the student is enrolled in.
- **KSU Campus_x**: Indicates whether the student belongs to KSU campus X
- **Stdng Stat**: The standing status of the student.to describe any action from honor code or other
- **Term_EOS:** The enrolled term details at the end of the semester
- **Descr_EOS:** The description of the enrolled term at the end of the semester.
- **Perm State_EOS:** The permanent location at the end of the semester.
- **Drop_out:** Indicates whether the student is considered dropped out or not. This the field to be predicted

The distribution of various features in the student data set is plotted using histograms (Barrett, et al., https://www.researchgate.net/publication/, 2005).

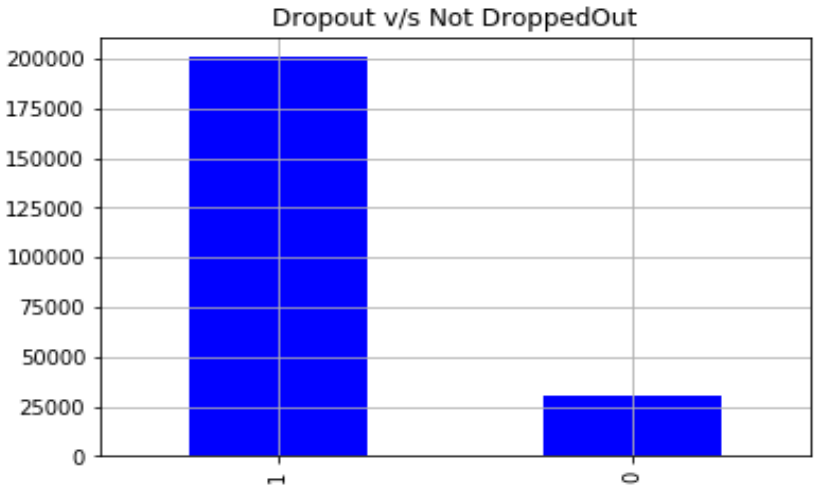


Figure 3.1 Attrition breakdown

The above histogram indicates the dropped-out student ratio versus the non-dropped out student count where 1 indicates the non-dropped out students and 0 indicates the dropped-out students.

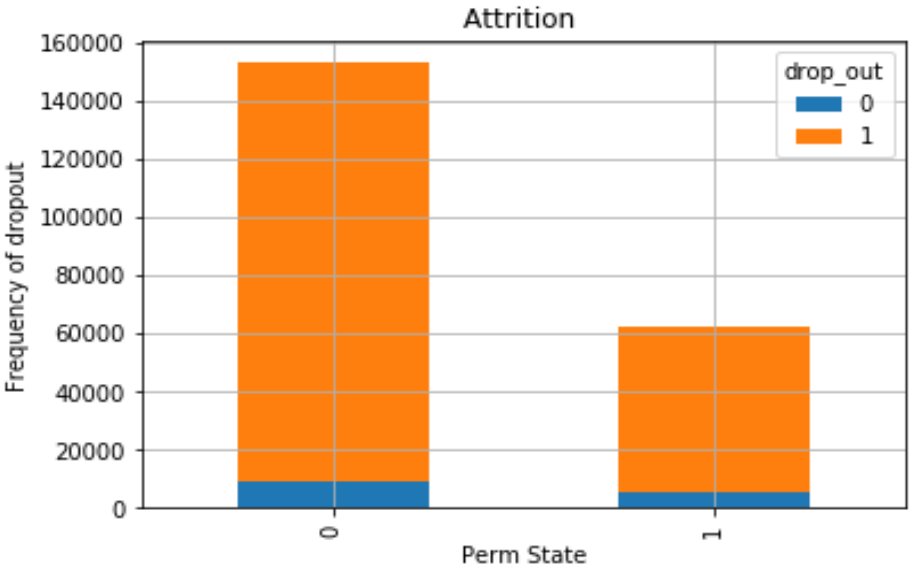


Figure 3.2 Attrition breakdown in Kansas

The above histogram shows the ratio of dropped out versus non dropped out in Kansas and outside the of Kansas state where 0 indicates the students from Kansas state and 1 indicates out of state.

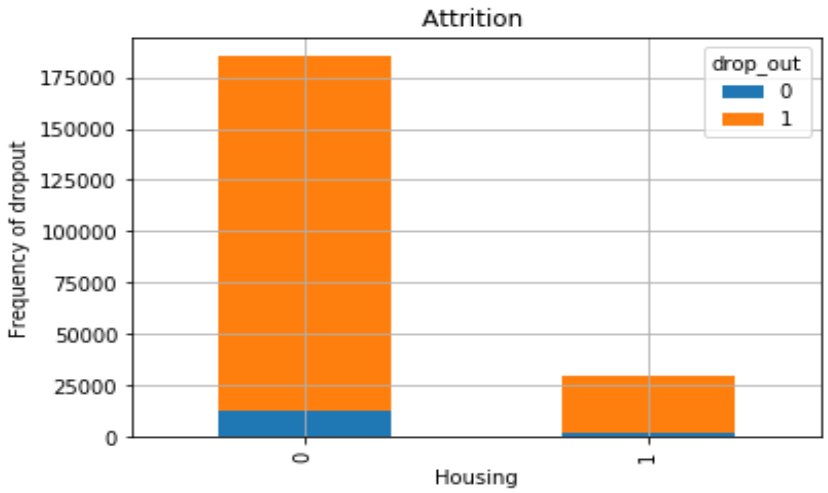


Figure 3.3 Attrition breakdown respective to housing

The above graph shows the ratio of dropped out versus non dropped out respective to the on campus and off campus students where 0 indicates the on-campus students and 1 indicates the off-campus students.

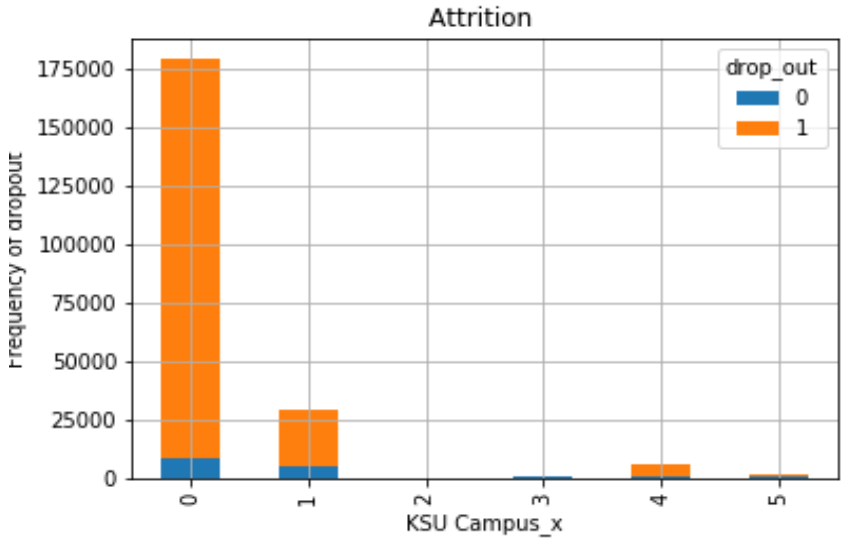


Figure 3.4 Attrition breakdown respective campus location

This graph is the ratio of dropped out versus non dropped out respective to the campus location.

3.2 Data Preparation

Data preparation is an important step in the machine learning process. This is the preparation of data to feed into the algorithm to improvise the results.

3.2.1 Data Preprocessing

The data preprocessing is a technique in data mining that helps in the transformation of raw data into a formatted form. The real-world data or the raw data consists of noise along with the signal. The noise in data will decrease the performance of the algorithm in predicting the dependent variable. The data entries that may establish the need for data preprocessing are following types

1. Missing values
2. Inconsistent entries
3. Noise

The missing values indicate the fields whose values are not filled. Inconsistent entries indicate that one type of data filed being filled with another type of data. Thus, the raw data should be preprocessed to eliminate the risk of the bad learning model. The Student data is processed and searched for the missing values. The few rows found with the missing values are handled properly by filling them with the values gathered from the end of the semester data and Graduation data. The Missing values are also handled by filling the missing value with dummy values like mean values.The programming language used in this project is python. The data preprocessing or cleaning, analysis and prediction tasks are carried out using the python library from the pandas.

The data is provided in the form of .csv files. The data is distributed into 3 files which consist of data at the beginning of the semester, the end of the semester and Graduation data. The data is then aggregated based on logical analysis and mapping. The aggregated data set is loaded into panda's data frame and divided into testing and training datasets using the libraries imported from the python sci-kit learn (Pedregosa, et al., n.d.). The algorithms for performing the prediction task on the student data set can be imported from the scikit learn.

```
Int64Index: 215183 entries, 0 to 317367
Data columns (total 16 columns):
Term                215183 non-null float64
Perm State          215183 non-null int64
Housing             215183 non-null int64
UGRD Admit Term     215183 non-null float64
GRAD Admit Term     215183 non-null float64
Readmit             215183 non-null int64
CUM GPA             215183 non-null float64
Semester Hours      215183 non-null float64
KSU Campus_x        215183 non-null int64
Stdng Stat          215183 non-null int64
CUM GPA_EOS         215183 non-null float64
Stdng Stat_EOS      215183 non-null int64
Term_DGA            215183 non-null float64
Descr_DGA           215183 non-null int64
diff_CGPA           215183 non-null float64
drop_out            215183 non-null int64
dtypes: float64(8), int64(8)
memory usage: 27.9 MB
None
```

Figure 3.5 Student data after handling the null values

It is also observed from the data set that the dependent variable categories are not equally distributed. The attrition ratio between the dropped out and non-dropped out students are represented in the following diagram

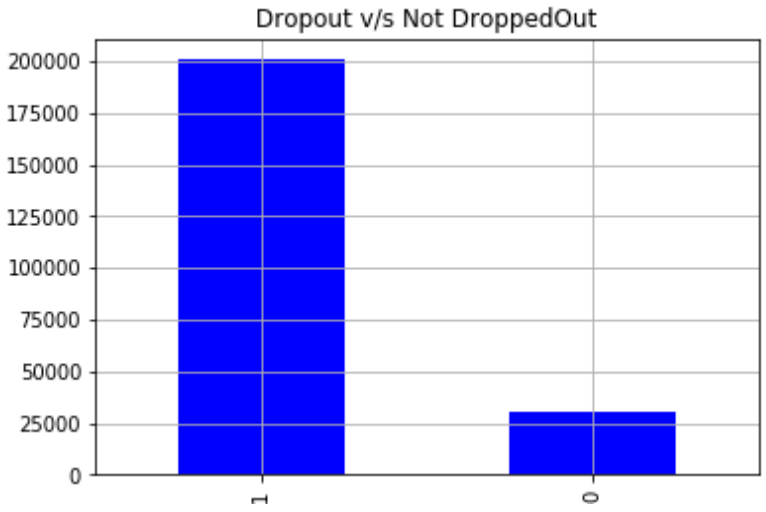


Figure 3.6 Attrition breakdown

From the above graph, it is very clear that the student ratio is not equally distributed between the dichotomous classes of the output. Students who did not drop out are comparatively in a higher ratio to the students who dropped out. This may prone the model risk of being partial to the non-dropped out class due to the presence of a larger set of training examples. To prevent this partiality in training the model a technique called SMOTE is applied. Synthetic Minority oversampling technique is developing of the artificial data sets for the minority classes. This process is developing similar data samples which are close to the existing data.

3.2.2 Feature Analysis

The feature analysis is an important task. The features which do not play any role in the prediction task or which does not help the model to learn or whose presence hinders the performance of the algorithm needs to be eliminated to improve the performance of the algorithm.

1. **Admit term:** This attribute the term or the semester in which the student is admitted to the program and the value of this is independent of predicting the output variable.
2. **Student ID Number:** This is the unique Identification given to the student whose value is different for each one and it does not play role in predicting the output.
3. **Permanent city:** This attribute value is simplified to a permanent state which helps to analyze whether a student is in state or Out of the state
4. **Permanent address:** This attribute value is also simplified and stored as the permanent state which is required in classification

Feature engineering is one of the important techniques that play a vital role in predicting the output. Various techniques are used to analyze the feature importance. Recursive feature elimination technique is used to estimate the importance of each feature using the co-eff_ and feature_importance_ of each feature. Depending on how feature importance is carried out the accuracy and precision of the model may increase or decrease. Regularization techniques like Lasso regression are also used to eliminate the features which play a negligible role in predicting the dependent variable.

3.2.3 Feature Importance Graph – Student Data

The features importance graph is plotted to estimate the importance of each feature on predicting the dependent variable

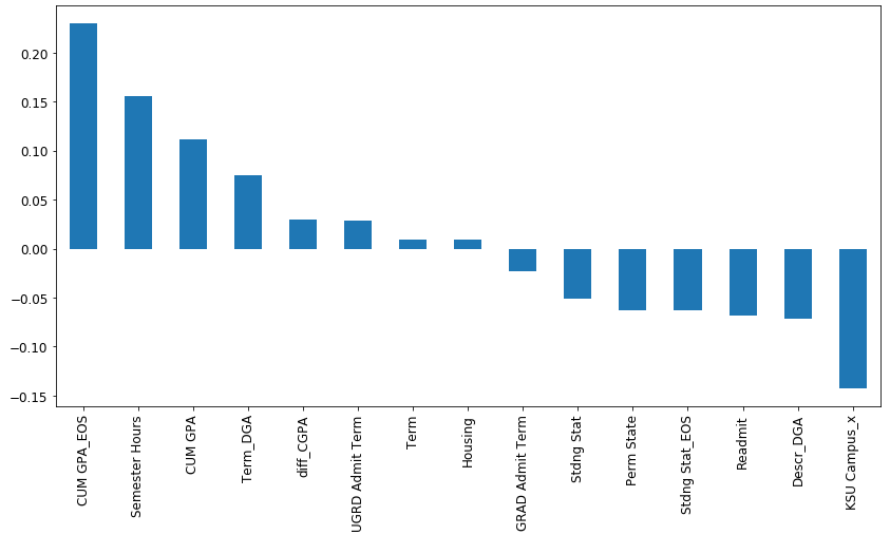


Figure 3.7 Coefficients of features the Regression models

We can see that coefficients of CUM GPA_EOS, Semester Hours, CUM GPA are higher as compared to the rest of the coefficients. Therefore, the drop out of s student would be more driven by these two features. The magnitude of coefficients in our model can be reduced using different types of regression techniques which uses regularization to overcome this problem.

3.2.4 Implementation Steps

The steps that are followed in the project to achieve the prediction as described below.

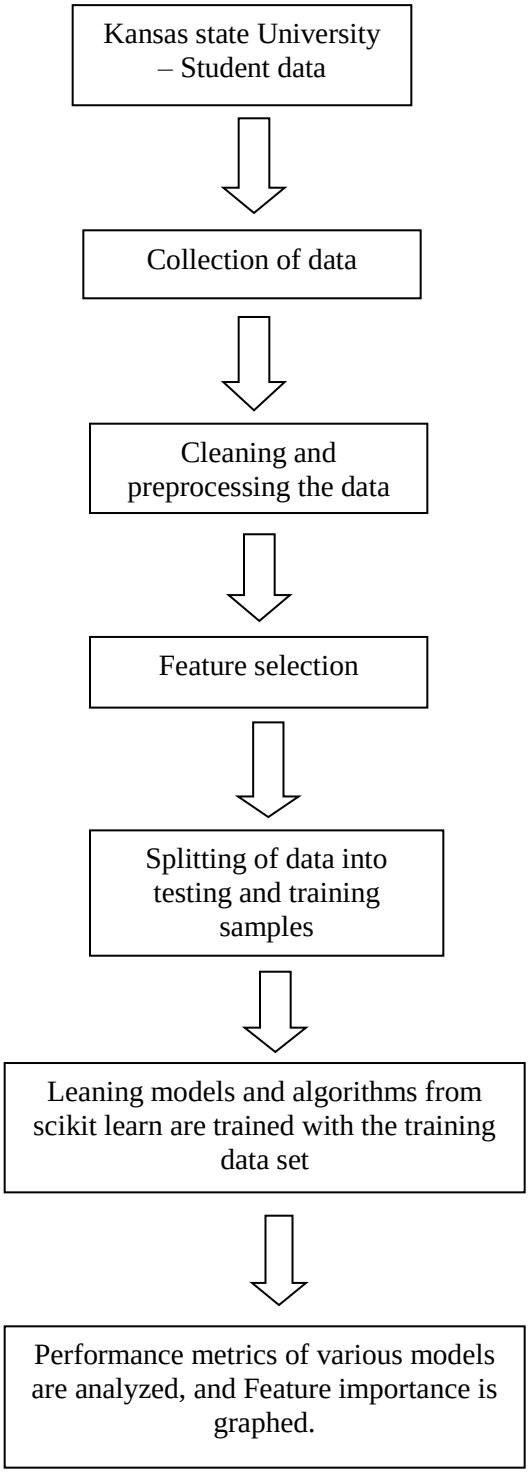


Figure 3.8 Process Flow of Project

The data is collected from Kansas state university Registrar office and includes 3 types of data. The data collected at the beginning of the semester, data collected at the end of the semester and data of the students graduating each semester. The data in the following files is merged to form a single meaningful data set eliminating the null values which is carried out during the data preprocessing stage. Then the features of data are analyzed, and the unwanted features are eliminated. The data set is further divided into training and testing sets using the packages from sci-kit learn and used in training various models for prediction of students drop out rate. The various metrics are used for the comparison of results and to determine the best algorithm for the prediction task.

Chapter 4 - Experiments

The series of experiments conducted on the data set are described below This chapter discusses a brief overview of this project, starting with describing the process of splitting data into testing and training sets followed by describing the procedure of conducting various experiments on the data.

4.1Training and Test Data Sets

The crucial task in training the machine learning models is to divide the data set between the training and the testing sets. The training set is used to train the algorithm so that the model learns from training set examples and helps to predict the test set data. The data for the following examples are divided into several ratios of test and train sets and tested for performance. The optimal performance was obtained when the data is divided as 70% and 30% where 70% is the training data set and 30% is the testing data set

Table 4.1 Student Data Set

Total records in the student data	215183
Training Records	150628
Testing Records	64555

4.2Experiment Design

The implementation design is that the above-discussed algorithms are trained and tested in the project. The algorithms are imported from the python library Scikit learn which generally has the prototype for all supervised and unsupervised machine learning algorithms. The data is split into testing and training sets using the test-train split in the scikit learn and later Testing sets are used for training the algorithm and the test sets are used for validating the algorithms.

Cross-validation is the technique of validating the model against the data set that was not used in training the model to eliminate the issues of overfitting and feature selection. It is the process of dividing the sample data into some subsets and performing the analysis on one subset and validating the model on another subset. Generally, the K-fold cross validation is one of the important and most used types of cross-validation. In this technique, the data set in randomly divided into k subsets. The model is trained with k-1 subsets and tested on the remaining 1 subset. This process is repeated k number of times each time considering a different subset for testing. Generally, 10-fold cross validation is preferred and is used in this project.

4.2.1 Logistic Regression

Logistic Regression is the supervised learning algorithm in machine learning that can be imported and used for validation in the following way: Initially, the Sklearn package is imported in the python and then the *Logistic Regression* is implemented by importing the *LogisticRegression* class *from sklearn. linear_model* (Li, n.d.). Then fit. predict_proba is used to fit the model on the training data set and *LogisticRegression.predict()* is used to test the model on the testing data set

4.2.2 Linear Regression

The Linear Regression algorithm was implemented by importing the library *from sklearn. linear_model import Linear Regression.* Then it fit against the training data set using *model. Fit* function and is testing against the validation data set using *model. Predict ()* function. The evaluation metrics like Mean squared error, mean absolute error and root mean squared error are imported *from sklearn import metrics* and are used to compare the performance.

4.2.3 Lasso Regression

The Lasso Regression algorithm was implemented by importing the library *from sklearn.linear_model import Lasso* (Pereira, Basto, & da Silva, 2016). Then the model is fit against the training data set using *model. fit* function and is tested against the validation data set using *model.predict()* function. The evaluation metrics like Mean squared error, mean absolute error and root mean squared error is imported *from sklearn import metrics* and are used to compare the performance. The testing, training scores and number of features used in training the model were also calculated using the functions in imported packages

4.2.4 Ridge Regression

The Ridge Regression is implemented by importing the library *from sklearn.linear_model import RidgeCV* (Pre). Then the model is fit against the training data set using *model. fit* function and is tested against the validation data set using *model.predict()*. The evaluation metrics like Mean squared error, mean absolute error and root mean squared error is imported *from sklearn import metrics* and are used to compare the performance. The testing, training scores and number of

features used in training the model were also calculated using the functions in imported packages

4.3 Evaluation Metrics

In the calculation of the evaluation metrics True Positives, True Negatives, False positives, and False negatives play a major role. They are explained below.

True Positives (TP): They are the correctly predicted positive values which mean the classes that are “yes” are predicted to be “yes”.

True Negatives (TN): They are the correctly predicted negative values which mean the classes that are “No” are predicted to be “No”

False Positives (FP): They are the incorrectly predicted positive values which determine classes that are “No” are predicted to be “yes”

False Negatives (FN): They are the incorrectly predicted Negative values which mean the classes which are “yes” are predicted to be “No”

4.3.1 Accuracy

Accuracy is the ratio of total correct predictions to the total number of observations. This is one of the best measures of accuracy only if we have the same number of false positives and false negatives.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

Equation 4.1 Accuracy Formula

4.3.2 Precision, Recall, and F-Score

Precision can be defined as the ratio of correctly predicted positive observations to the total number predicted observations

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

Equation 4.2 Precision Formula

Recall can be defined as the ratio of correctly predicted positive observations to the total number of observations present in the class – Positive

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

Equation 4.3 Recall Formula

F1 Score is the harmonic mean of precision and recall. Usually, F1 score is a better measure than accuracy when we have the unevenly distributed classes

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{2 * TP}}{\boldsymbol{2 * TP + FP + FN}}$$

Equation 4.4 F1 - Score Formula

4.3.3 Receiver Operating Characteristic Curve (ROC)

The ROC curve is one of the important metrics in analyzing the performance of the model. It is plotting of the true positive rate against false positive rates at various threshold settings. The true positive rate is defined as precision and the false positive rate is defined as 1-specificity.

The area under the curve (AUC) is the probability that the model will rank the randomly chosen positive value higher than the randomly chosen negative instance. The maximum value for the area under the curve is 1 and the minimum of 0.5 is considered as a good model. The higher the AUC value the model performs better. *roc_auc_score* and *roc_curve* are the metrics used in measuring the performance and are imported by *from sklearn.metrics import roc_curve, roc_auc_score.*

4.3.4 Cross-validation

The K-fold cross validation is one of the important and most used types of cross-validation. In this technique, the data set in randomly divided into k subsets. The model is trained with k-1 subsets and tested on the remaining 1 subset. This process is repeated k number of times each time considering a different subset for testing. Generally, 10-fold cross validation is preferred and is used in this project. This is implemented by importing *cross_val_score* function from *sklearn.model_selection*.

4.3.5 Confusion Matrix

A *confusion matrix* is a tabular representation of the algorithm performance on the testing data set. The rows of the matrix represent the predicted class whereas the columns represent the

actual values or the actual classes. This matrix is a useful representation to identify the confusion of the algorithm in predicting the values. The sum of all the values in the confusion matrix is the total number of records in the testing set. The smaller is the sum of values in the diagonal starting from the right top corner to the left bottom corner the better is the algorithm performance.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

(Walia, 2019)

Figure 4.1 Confusion Matrix

4.3.6 Regression Metrics:

The important task in comparing the metrics of various algorithms is choosing the metrics of comparison (Drakos, G., & Drakos, n.d.). The Mean absolute error, Root mean squared error, mean squared error are the important metrics chosen to evaluate the performance of regression algorithms.

Mean Squared Error:

This metric calculates the average sum of squares of the difference between the actual and predicted values. The value is never going to be negative, but the perfect model will have the value 0. The higher the value is, the less effective the model will be.

$$MSE = \frac{1}{N}\sum_{I=1}^{N}(y_i - \hat{y_i})^2$$

Equation 4.5 Mean squared error formula

Root Mean Squared Error:

This metric is calculated by performing the square root of the mean squared error. In order to make the scale of errors equal to the scale of the target, the square root of the value is performed.

$$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(y_i - \hat{y_i})^2}$$

Equation 4.6 Root Mean squared error formula

Mean Absolute Error:

The mean absolute error is the average of the absolute differences between the actual and the predicted values. This MAE can penalize huge error but not as good as Mean squared error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}|$$

Equation 4.7 Mean absolute error formula

Chapter 5 - Results

This chapter includes the results of the experiments mentioned in Section 4.2.1 - 4.3.6 and the evaluation metrics used to evaluate the performance of those machine learning models.

5.1 Experimental Results for Student Data

The results obtained for the algorithms discussed in Chapter 4 are explained here using their performance metrics.

5.1.1 Student Data – Logistic Regression:

The following is the classification report and comparison of metrics like the Roc curve, Confusion matrix, etc. when the logistic regression is run on the student data.

Table 5.1 Classification Report of Logistic Regression

	Precision	**Recall**	**F1- Score**	**Support**
0	0.78	0.12	0.21	4777
1	0.94	1.00	0.97	66234
Micro avg	0.94	0.94	0.94	71011
Macro avg	0.86	0.56	0.59	71011
Weighted avg	0.93	0.94	0.92	71011

Table 5.2 Evaluation Metrics - Logistic Regression

Evaluation Metric	Value
Precision	0.124
Recall	0.997
F1 score	0.22
Area under ROC Curve	0.560

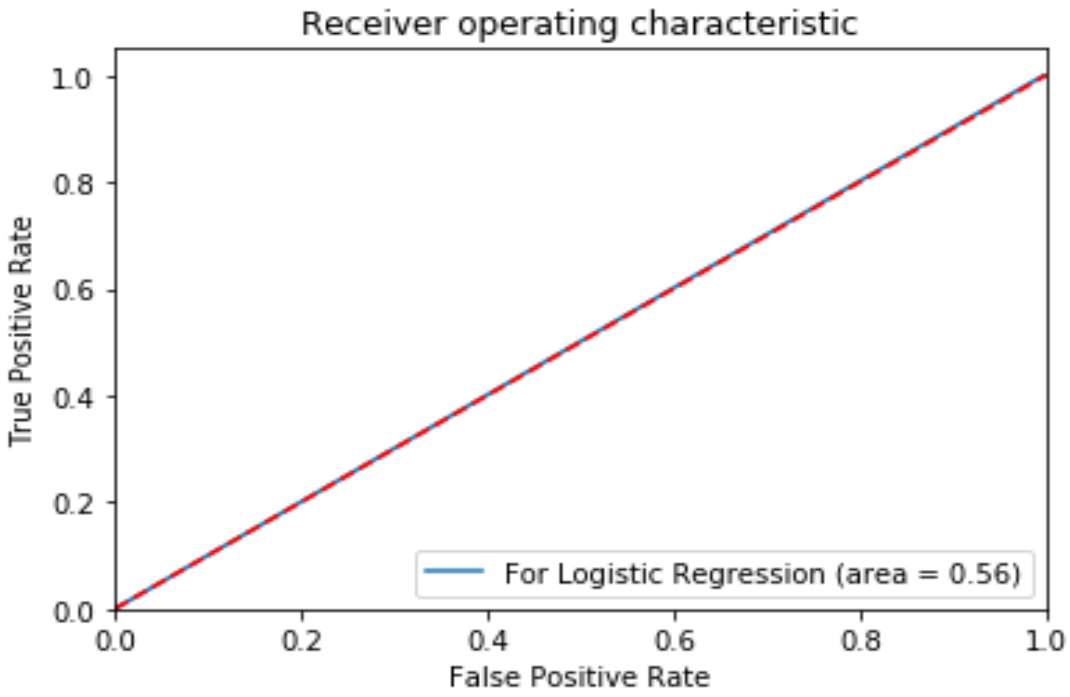


Figure 5.1 ROC Curve – Logistic Regression

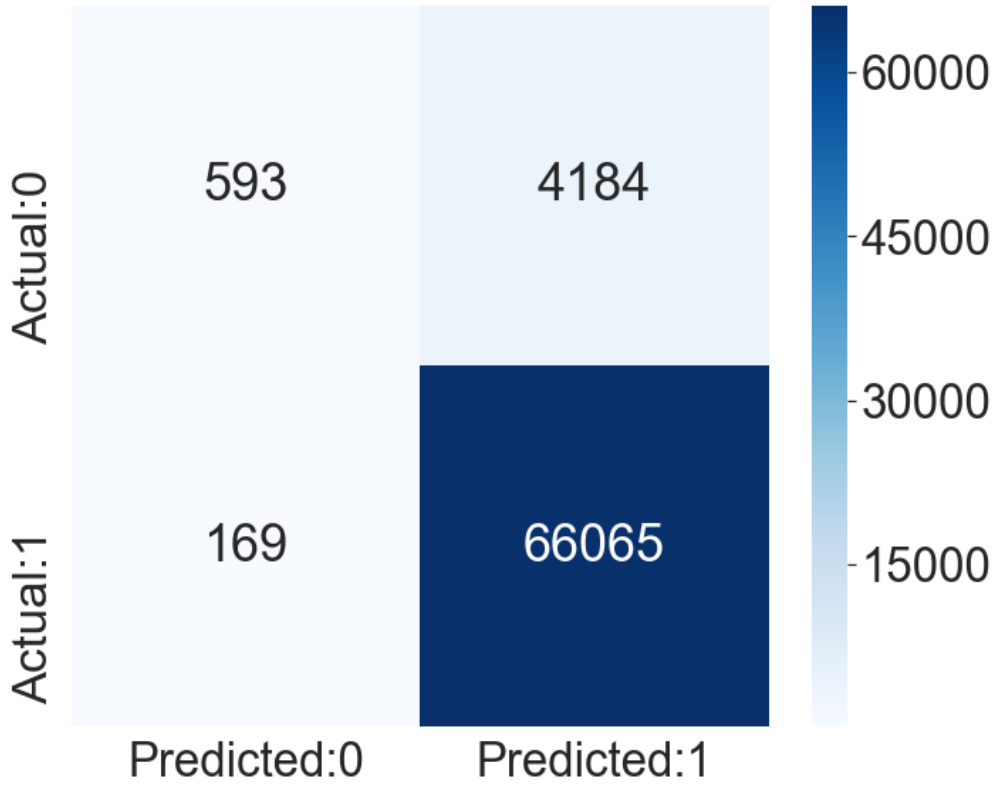


Figure 5.2 Confusion Matrix for Logistic Regression

5.1.2 Student Data – Linear Regression:

The following are the coefficients of various features, differences between various coefficients and the evaluation metrics for the regression models like RMSE, MAE, MSE when the linear regression is run on the student data.

```
**********Coefficients of various features**********
CUM GPA_EOS.............  2.5126e-01
diff_CGPA...............  2.1562e-01
CUM GPA................. -2.0561e-01
Readmit................. -9.4970e-02
Perm State.............. -3.0216e-02
KSU Campus_x............ -2.6761e-02
Housing................. -1.3484e-02
Stdng Stat_EOS.......... -1.3342e-02
Stdng Stat..............  8.0035e-03
Semester Hours..........  7.2423e-03
Term_DGA................  4.5220e-03
Term....................  4.9629e-04
Descr_DGA............... -2.8892e-04
GRAD Admit Term.........  7.7403e-06
UGRD Admit Term.........  1.6004e-06
```

Figure 5.3 Coefficients of various features in Linear Regression

Table 5.3 Regression Evaluation Metrics - Linear Regression

Evaluation Metric	Value
Mean Squared Error	0.055
Root Mean squared error	0.235
Mean Absolute Error	0.116

Table 5.4 Training and Testing scores - Linear Regression

Evaluation Metric	Value
Training Score Error	0.134
Testing score	0.142
Number of features used	15

5.1.3 Student Data – Linear Regression with L1 Regularization:

The following are the coefficients of various features, differences between various coefficients and the evaluation metrics for the regression models like RMSE, MAE, MSE when the L1 Regularization is applied on model fitted with Linear Regression on the student data.

Table 5.5 Classification Report of Linear Regression - L1 Regularization

	Precision	Recall	F1- Score	Support
0	1.00	0.07	0.14	5957
1	0.94	1.00	0.97	80117
Micro avg	0.94	0.94	0.94	86074
Macro avg	0.97	0.54	0.55	86074
Weighted avg	0.94	0.94	0.91	86074

Table 5.6 Evaluation Metrics for Linear Regression - L1 Regularization

Evaluation Metric	Value
Precision	0.068
Recall	1.0
F1 score	0.127
Area under ROC Curve	0.769

Table 5.7 Regression Evaluation Metrics – Linear Regression - L1 Regularization

Evaluation Metric	Value
Mean Squared Error	0.055
Root Mean squared error	0.235
Mean Absolute Error	0.116

Table 5.8 Training and Testing scores - Linear Regression - L1 Regularization

Evaluation Metric	Value
Training Score Error (alpha= 0.00001)	0.137
Testing score (alpha = 0.00001)	0.134
Training Score Error (alpha= 0.01)	0.090
Testing score (alpha=0.01)	0.091
Number of features used	15

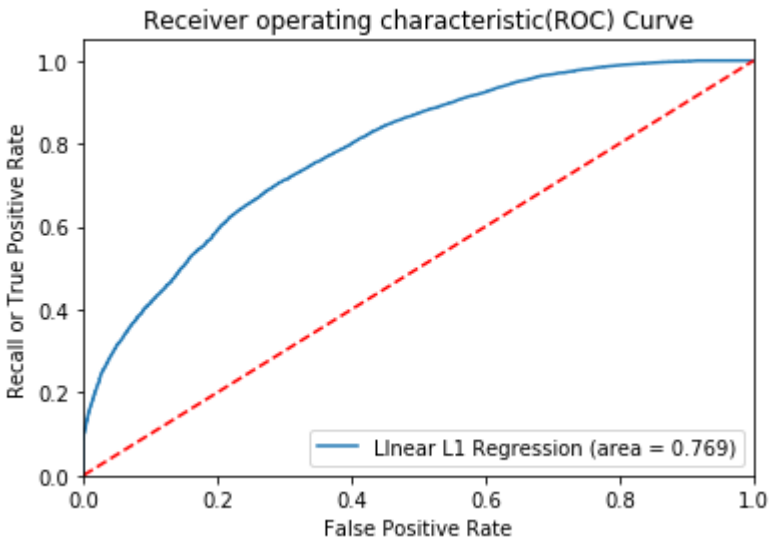


Figure 5.4 ROC Curve – Lasso Regression

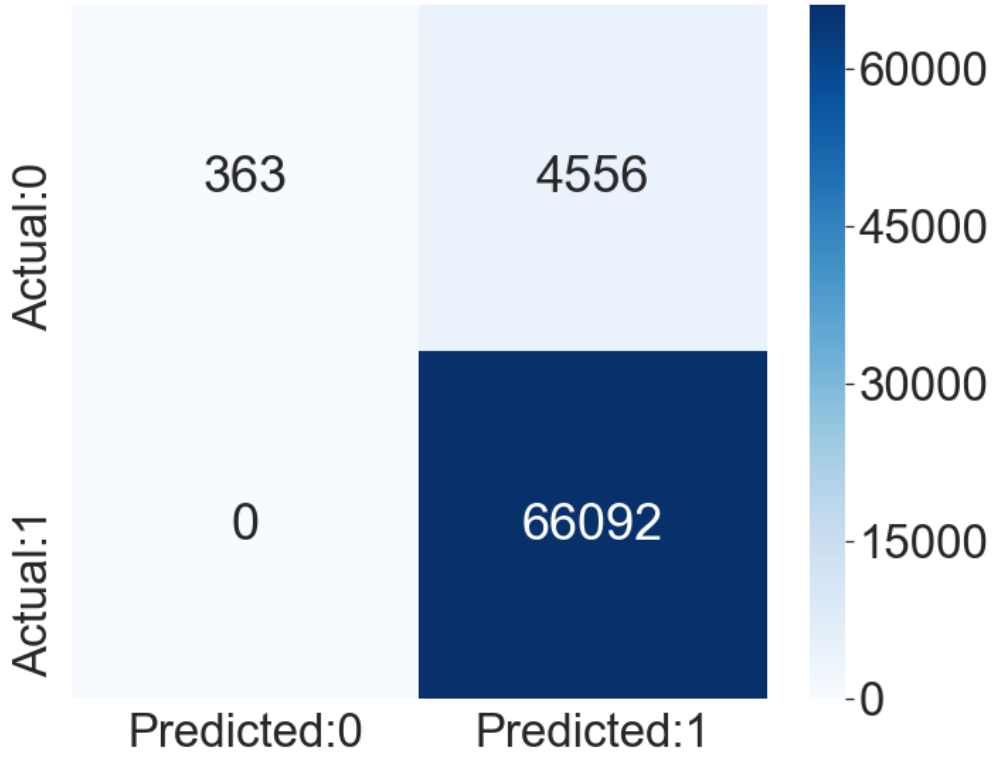


Figure 5.5 Confusion Matrix – Linear Regression - L1 Regularization

5.1.4 Student Data – Linear Regression with L2 Regularization

The following are the coefficients of various features, differences between various coefficients and the evaluation metrics for the regression models like RMSE, MAE, MSE when the L2 Regularization is applied on model fitted with Linear Regression on the student data.

Table 5.9 Classification Report of Linear Regression – L2 Regularization

	Precision	**Recall**	**F1- Score**	**Support**
0	1.00	0.07	0.14	5957
1	0.94	1.00	0.97	80117
Micro avg	0.94	0.94	0.94	86074
Macro avg	0.97	0.54	0.55	86074
Weighted avg	0.94	0.94	0.91	86074

Table 5.10 Evaluation Metrics – Linear Regression – L2 Regularization

Evaluation Metric	**Value**
Precision	0.073
Recall	1.0
F1 score	0.136
Area under ROC Curve	0.78

Table 5.11 Regression Evaluation Metrics - Student Data

Evaluation Metric	Value
Mean Squared Error	0.055
Root Mean squared error	0.235
Mean Absolute Error	0.116

Table 5.12 Training and Testing scores - Linear Regression – L2 Regularization

Evaluation Metric	Value
Training Score Error (low alpha)	0.134
Testing score (low alpha)	0.142
Training Score Error (High alpha)	0.134
Testing score (High alpha)	0.142
Number of features used	15

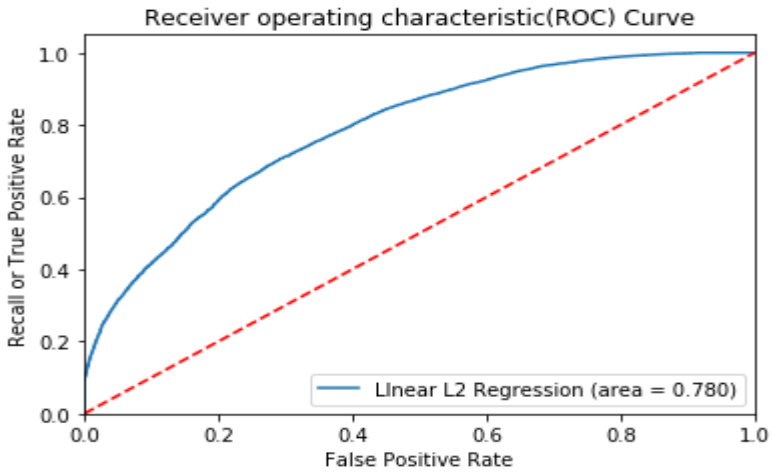


Figure 5.6 ROC Curve - Ridge Regression

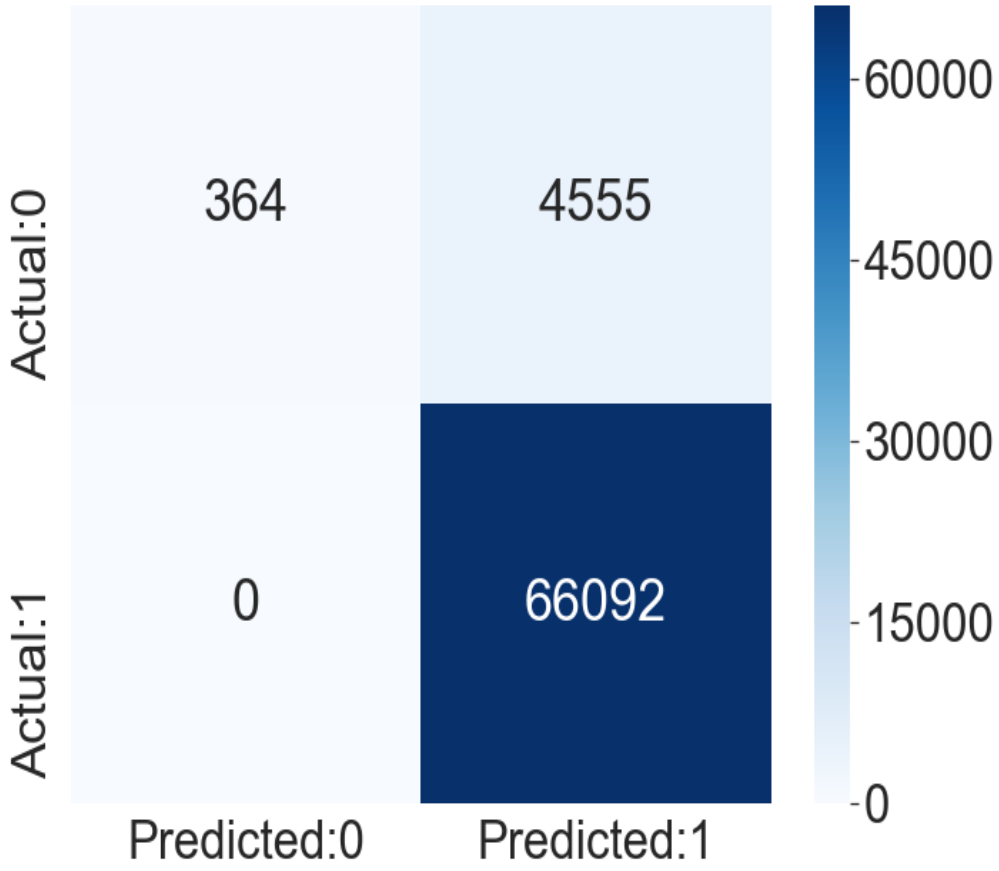


Figure 5.7 Confusion Matrix – Linear Regression – L2 Regularization

5.1.3 Student Data – Logistic Regression with L1 Regularization

The following are the coefficients of various features, differences between various coefficients and the evaluation metrics for the regression models like RMSE, MAE, MSE when the L1 Regularization is applied on model fitted with Logistic Regression on the student data.

Table 5.13 Classification Report of Logistic Regression – L1 Regularization

	Precision	Recall	F1- Score	Support
0	1.00	0.07	0.14	5957
1	0.94	1.00	0.97	80117
Micro avg	0.94	0.94	0.94	86074
Macro avg	0.97	0.54	0.55	86074
Weighted avg	0.94	0.94	0.91	86074

Table 5.14 Evaluation Metrics – Logistic Regression – L1 Regularization

Evaluation Metric	Value
Precision	0.077
Recall	1.0
F1 score	0.142
Area under ROC Curve	0.774

Table 5.15 Regression Evaluation Metrics - Student Data

Evaluation Metric	Value
Mean Squared Error	0.055
Root Mean squared error	0.235
Mean Absolute Error	0.116

Table 5.16 Training and Testing scores - Logistic Regression – L1 Regularization

Evaluation Metric	Value
Training Score Error (low alpha)	0.134
Testing score (low alpha)	0.142
Training Score Error (High alpha)	0.134
Testing score (High alpha)	0.142
Number of features used	15

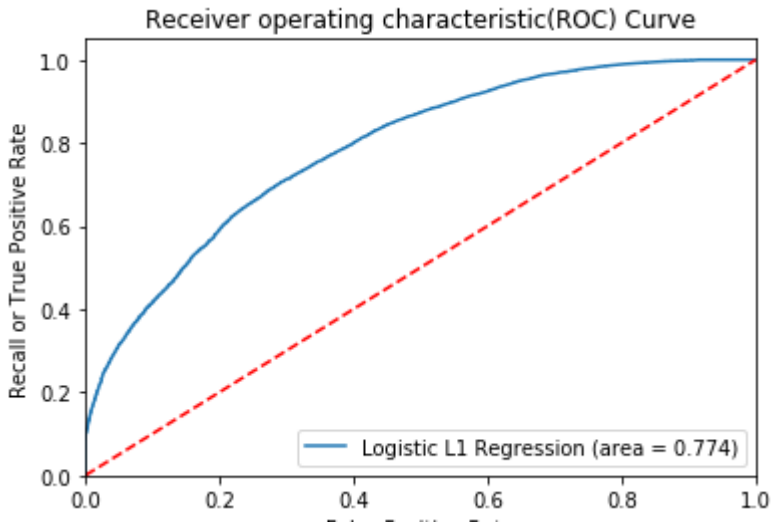


Figure 5.8 ROC Curve - Lasso Regression

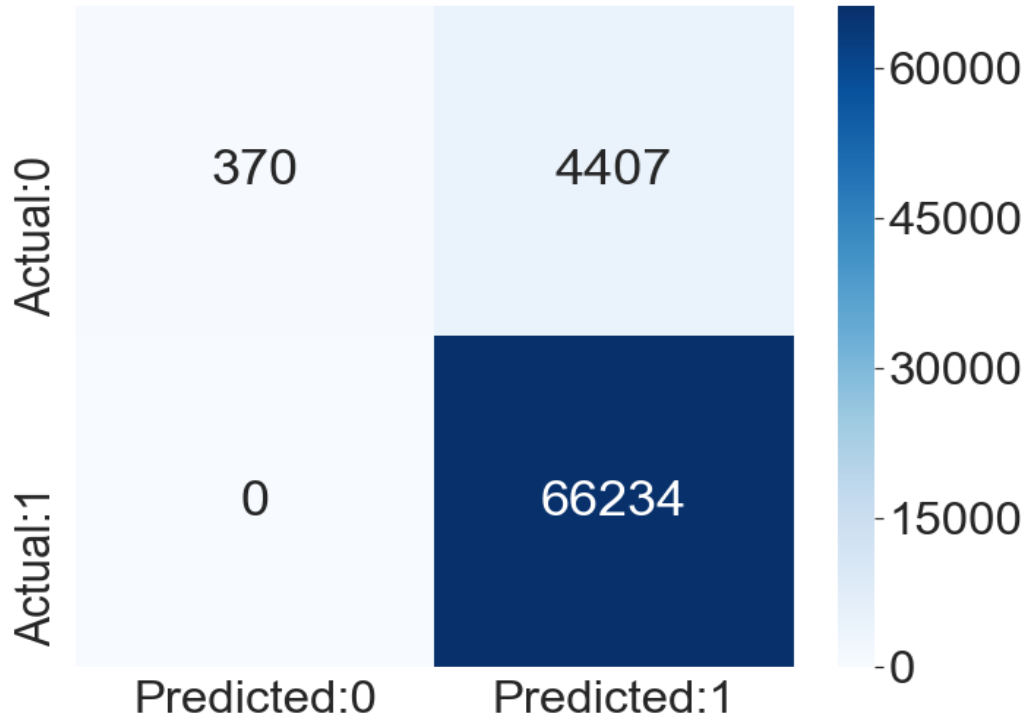


Figure 5.9 Confusion Matrix - Logistic Regression – L1 Regularization

5.1.6 Student Data – Logistic Regression with L2 Regularization

The following is the coefficients of various features, differences between various coefficients and the evaluation metrics for the regression models like RMSE, MAE, MSE when the L2 Regularization is applied on model fitted with Logistic Regression on the student data.

Table 5.17 Classification Report of Logistic - Ridge Regression

	Precision	Recall	F1- Score	Support
0	1.00	0.07	0.14	5957
1	0.94	1.00	0.97	80117
Micro avg	0.94	0.94	0.94	86074
Macro avg	0.97	0.54	0.55	86074
Weighted avg	0.94	0.94	0.91	86074

Table 5.18 Evaluation Metrics – Logistic - Ridge Regression

Evaluation Metric	Value
Precision	0.073
Recall	1.0
F1 score	0.136
Area under ROC Curve	0.78

Table 5.19 Regression Evaluation Metrics - Student Data

Evaluation Metric	Value
Mean Squared Error	0.055
Root Mean squared error	0.235
Mean Absolute Error	0.116

Table 5.20 Training and Testing scores - Ridge Regression

Evaluation Metric	Value
Training Score Error (low alpha)	0.134
Testing score (low alpha)	0.142
Training Score Error (High alpha)	0.134
Testing score (High alpha)	0.142
Number of features used	15

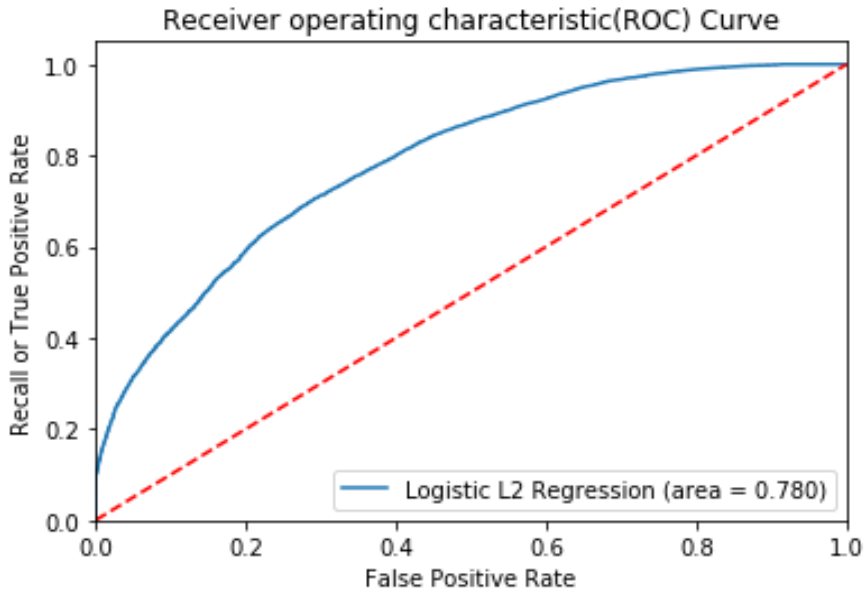


Figure 5.10 ROC Curve - Ridge Regression

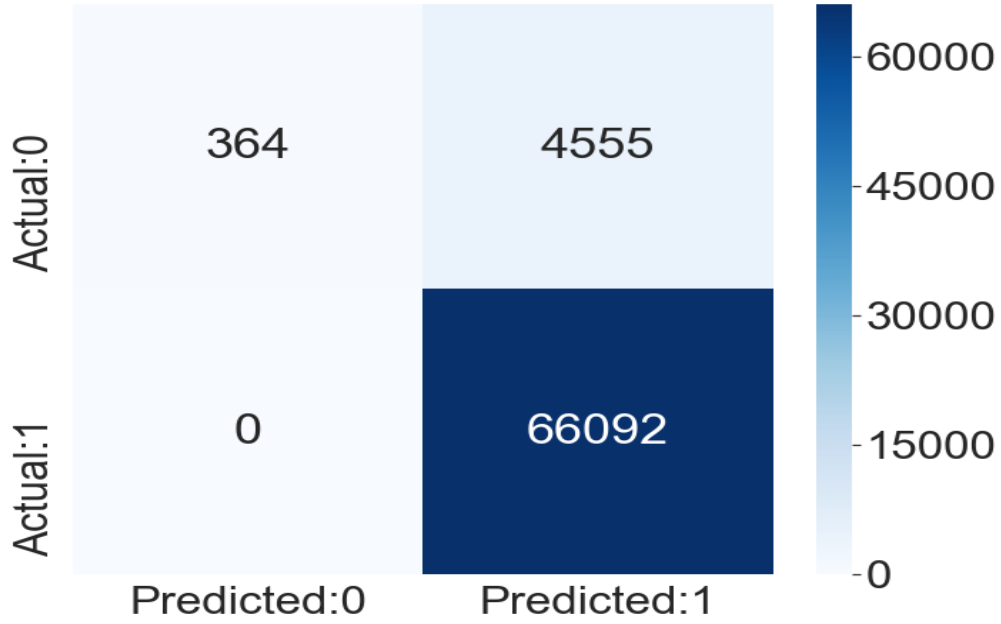


Figure 5.11 Confusion Matrix - Ridge Regression

5.2 Comparison of all Classifiers

5.2.1 Precision, recall, F1 score values

This Section presents the performance results of classifiers at the task of predicting student drop out. The precision Recall and F1- Score are only considered since the accuracy does not make much sense due to class Imbalance. The input data has more samples under the positive case i.e. in which the students did not drop out of university which will create partiality on True Positives detection accuracy. It is seen that the L1 Regularization technique on Logistic Regression gave a Precision of 0.077 and a recall value of 1.00 which are considered to higher compared to all other models implemented. It also showed a high F1- score of 0.142 compared to other models.

Table 5.21 Precision, recall, F1 Score values for all classifiers

Classification model	**Precision**	**Recall**	**F1-Score**
Logistic Regression with RFE	0.124	0.997	0.22
Logistic Regression without RFE	0.124	0.997	0.22
Linear Regression – L1 Regularization	0.068	1	1.127
Linear Regression - L2 Regularization	0.073	1	0.136
Logistic Regression - L1 Regularization	**0.077**	**1**	**0.142**
Logistic Regression - L2 Regularization	0.073	1	0.136

5.2.2 The Area Under the ROC Curve

This Section presents the Area under Curve for ROC Curve for the classifiers at the task of predicting student drop out. It is seen that the L2 Regularization technique on Logistic Regression and Linear Regression gave a value of 0.78 for the area under the curve which is considered high when compared to all other models implemented. These models which showed high AUC values showed a precision of 0.073 and a recall value of 1.00 which were second better values when compared to all other models and L1 Regularization on Logistic Regression showed the better precision and recall values compared to remaining models.

Table 5.22 Area under ROC curve for all classifiers

Classification model	**The area under the curve**
Logistic Regression with RFE	0.56
Logistic Regression without RFE	0.56
Linear Regression – L1 Regularization	0.769
Linear Regression - L2 Regularization	**0.78**
Logistic Regression - L1 Regularization	0.774
Logistic Regression - L2 Regularization	**0.78**

Chapter 6 - Summary and Future Scope

6.1 Summary and Interpretation of Results

Based on the experimental results, it was observed that the algorithms outperformed the results after performing the regularization techniques on the Logistic and Linear regression models. Also, the comparison metrics seem to improve when the size of the testing data set is increased. The Logistic Regression with the L1 Regularization technique model outperformed the remaining models showing a recall of value 1.00 and precision values of 0.077 which is considered extremely beneficial in building a model with 0 false negatives. The L2 regularization technique applied to Linear and Logistic Regression showed better AUC values compared to all the remaining models. The Regularization techniques helped in feature selection and solving the issue of overfitting thereby improving the results compared to the baseline models. Also, the algorithms showed good results when low alpha values were used in Regularization techniques.

6.2 Future Scope

6.2.1 Domain Specific

The data set used for the analysis and prediction task is obtained in a very raw format. The data consisted of many missing values and many features are not playing any major role in predicting the output which is eliminated during the feature selection. So, a data set with more features that play the major role in predicting the student dropout rate like Gender, Financial status, working hours per week, pay rate, off campus or on campus working category needs to be collected from the registrar office. Feedback should be collected from the dropped-out students requesting for the reasons to drop out which would contribute a lot to the prediction task.

6.2.2 Methodology of Applying Machine Learning Techniques

. Bagging attempts to reduce the overfitting and the boosting methods to improve predictive flexibility from multiple models can be applied to the data and the algorithms that ignore the missing values can be used in prediction tasks. Regularization techniques can be applied to other algorithms like Random Forest, Support vector machines, Decision Trees, etc. The comparison metrics for these algorithms upon applying regularization techniques like L1, L2 regularization can be compared to improve the results. Various techniques to solve the problem of class Imbalances like resampling the data with different ratios apart from down sampling and oversampling can be applied and results can be analyzed.

Chapter 7 - References

Bani, M. J., & Haji, M. (2017). College Student Retention: When Do We Losing Them? *arXiv preprint arXiv:1707.06210*.

Barrett, P., Hunter, J., Miller, J. T., Hsu, J. C., & Greenfield, P. (2005, December). matplotlib--A Portable Python Plotting Package. In *Astronomical data analysis software and systems XIV* (Vol. 347, p. 91).

Bonham, L. A., & Luckie, J. A. I. (1993). Taking a break in schooling: Why community college students stop out. *Community College Journal of Research and Practice*, *17*(3), 257-270.

DesJardins, S. L., & McCall, B. P. (2010). Simulating the effects of financial aid packages on college student stopout, reenrollment spells, and graduation chances. *The Review of Higher Education*, *33*(4), 513-541.

Johnson, N. (2012). The Institutional Costs of Student Attrition. Research Paper. *Delta Cost Project at American Institutes for Research*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

"Predict House Prices with Ridge Regression." *Kaggle* (2019). www.kaggle.com/miguelrodriguezolmos/predict-house-prices-with-ridge-regression

Pereira, J. M., Basto, M., & da Silva, A. F. (2016). The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*, *39*, 634-641.

Saptashwa. "Ridge and Lasso Regression: A Complete Guide with Python Scikit-Learn." *Towards Data Science*, Towards Data Science, 26 Sept. 2018, towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.

Predict House Prices With Ridge Regression. (n.d.). Retrieved April, 2019, from https://www.kaggle.com/miguelrodriguezolmos/predict-house-prices-with-ridge-regression

Bonham, L. A., & Luckie, J. A. I. (1993). Taking a break in schooling: Why community college students stop out. *Community College Journal of Research and Practice*, *17*(3), 257-270.

Drakos, G., & Drakos, G. (2018, August 26). How to select the Right Evaluation Metric for Machine Learning Models: Part 1 Regression Metrics. Retrieved from https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0