From High School to College: Exploring the Relationship between Secondary School performance and University Graduation

“University of Nebraska at Omaha”

“Bikram Maharjan”

12/12/2022

Abstract

The transition from high school to college can be a challenging experience for many students, with academic expectation and demands significantly increasing. As such, exploring the relationship between high school performance and college graduation rates is essential in understanding the factors that contribute to student success in college. This paper examines the relationship between high school performances and college graduation rates among students and identifies the factors that influence these outcomes. By understanding the relationship between these factors we can be prepared and educate new generation of students for success beyond high school.

1. Introduction

The relationship between High School performance and college graduation has been a topic of interest to educators and researchers for decades (citation). Many studies have explored various factors that contribute to academic outcomes in both high school and college, including factors like socioeconomic status, academic preparation, and personal characteristics such as motivation and perseverance (citation). But the role of standardized test scores and other factors such as school attendance, demographic, socioeconomic background, high school Zip Code etc in predicting college success has been a subject of continued debate (citation)(GPA is better).

In recent decades, Nebraska has witnessed the implementation of couple of different standardized tests. The statewide standardized test in Nebraska from the academic years 2009-2010 to academic year 2016-2017 was called the Nebraska State Accountability (NeSA), however, starting from the academic year 2017-2018, the Nebraska Student-Centered Assessment System (NSCAS) took over as the replacement for NeSA. Currently, the NSCAS serves as the standardized test used to assess student performance and provide accountability for schools and districts in the state of Nebraska. Both of these standardized tests administered to students in Nebraska is to measure their proficiency in core academic subjects and are intended to help educators and policymakers identify areas of improvement in student achievement.

In this paper, we seek to explore the relationship between high school performance and college graduation using factors such as but not limited to standardized test, demographic status, gender, homelessness, school attendence, high school zip code, etcs in shaping this relationship and delve into our findings. These factors pertain to the high school level of education. By analyzing these findings, our objective is to contribute to a more profound understanding of the factors that exert influence on college success. Furthermore, we also aim to provide valuable insights that can guide policymaking and inform practical implementation in the State of Nebraska and beyond.

1.1 Motivation

Education is widely recognized as a key determinant of economic and social mobility, and college completion rates are an important indicator of success in this regard (citation). High school performance has been 1 identified as a key predictor of college graduation outcome, with students who perform well academically in high school are more likely to enroll in and complete college degrees. There could be several factors that may influence the outcome, some of which we already know and some we still need to discover. Previous studies have consistently shown that students who perform well in high school are more likely to successfully complete a university degree . However, the relationship between High School performance and college success is a complex equation and is influenced by a range of factors. This relationship is a critical area of research within the field of education.

The importance of researching and understanding the impact of high school performance on college outcomes is paramount. Understanding the factors that influence student success during the critical transition can pave the way for improved educational practices and policies. By delving into this area of study, we gain valuable insights into the strengths and weaknesses of our education systems, enabling us to identify effective strategies for enhancing college readiness for students and hence have successful completion rates. Ultimately, this research empowers us to empower students, equipping them with the necessary tools to overcome obstacles and achieve their full potential in higher education and beyond.

Our curiosity drives us to explore the influence of various factors, including but not limited to standardized tests, school attendance, and high school locality, on students' academic progress. Our ultimate goal is to contribute to the success of students in their educational journey and beyond. By utilizing data-driven approaches, we seek to uncover and understand the underlying factors that shape these outcomes, empowering us to make informed decisions and implement effective strategies to support student achievement.

1.2 Significance of the work

This research and data analysis holds significant importance as its outcomes have the potential to positively impact the future of younger generations. It could help identify specific factors that contribute to academic success, allowing educators and policymakers to design improved ways of learning.

Additionally, our research aims to extend the existing body of knowledge in this field by incorporating additional factors beyond the commonly studied variables such as standardized tests or ethnicity. By considering factors such as high school attendance, locality (e.g., zip code), and demographic characteristics, our study offers a novel perspective that provides a more comprehensive understanding of the dynamics impacting educational outcomes. This unique approach enables us to explore previously unexplored dimensions and uncover potential insights that can inform policy decisions within the Nebraska Department of Education. By utilizing data-driven analyses, we aim to equip policymakers with evidence-based information, facilitating informed decision-making and ultimately contributing to the improvement of educational practices and outcomes in the region that will benefit students, educators, employers, and policymakers alike.

1.3 Research Objective

The goal of the research is understanding two key aspects:

1. Does high school performance impacts college outcomes?

   - We aim to investigate the connection between high school performance and subsequent college success.
   - By examining various factors and analyzing the relationship between these two educational stages, our study seeks to understand the extent to which performance in high school predicts or influences performance in college.

2. What are the factors that influence similar outcomes?

   - Here, we aim to examine the influential factors that contribute to the outcomes. Our goal is to determine specific areas of improvements that could require added focus in the secondary school.

1.4 Research Focus

As part of the research, our focus was solely on students who took the NeSA standardized test and was part of graduating high school class of 2012, 2013, 2014 due to the limited availability of data pertaining to students who took the NSCAS standardized test. This limitation arises from the fact that the NSCAS standardized test was introduced only in the academic year 2017-2018. Considering the typical timing of the NSCAS test administration during a student's 11th grade, the expected graduation year for the first cohort students entering a 4-year college would be around the 2024-2025. Hence, we chose to work with NeSA test scores only.

This research also specifically focuses on the students who pursued a 4-year college education, excluding those who attended either a 2-year institution or pursued a combination of 2-year and 4-year education paths. For the purpose of our study, we made an assumption that the maximum expected time for students enrolled in a 4-year college going graduation would be six years (1.5 * degree length or 1.5 * 4).

In summary, we will focus on the students who took NeSA standardized test and graduated high school in the year 2012, 2013, 2014, and then continued their education journey towards a 4-year college only. < what are the main areas of investigation or the central themes that the research study aims to address? >

2. Working with Data

Data is central to our research, serving as the foundation for our analysis and insights. By leveraging comprehensive datasets, we can systematically explore variables, patterns, and relationships to address our two important research objectives descrived above. This data-driven approach ensures that our findings are grounded in objective evidence and rigorous analysis, fostering the credibility, validity, robustness, and trustworthiness of our research outcomes.

2.1 Aggrement and Compliance with Nebraska Department of Education

The research work was conducted in collaboration with the Nebraska Department of Education (NDE) who graciously facilitated the project, provided us the platform, and the dataset encompassing both high school and college-level information for the research.

In adherence to our commitment to data privacy and confidentiality, a non-disclosure agreement was established with NDE. The research activities, including data analysis and the development of research outcomes, were carried out on a secured remote Azure Windows Platform server provided by NDE. This ensured the protection of sensitive data and maintained the integrity and privacy of the research process.

2.2 Data Overview

High School dataset

High school data contained individual student performance data per grade level per academic years. This encompassed NeSA and NSCAS scores, attendance records, and demographic characteristics (e.g., gender, homelessness, race/ethnicity, and immigrant status among others; refer to Appendix A for the complete list). The dataset encompasses information from the graduating classes from the year 2007 to 2016, spanning across approximately 80 public and private schools in Nebraska.

College dataset

College level dataset contained information post high school studies. This dataset encompassed a range of factors including the colleges/universities attended by each student, degree titles, majors, states graduation 3 status, graduation dates and detailed semester level data such as the start and end dates of each semester, enrollment status (e.g., full-time, part-time, etc.), and more.

Table 1: Variables used from each dataset

High.School	College
StudentID, Datayears, AgencyID, DistrictName, SchoolName, GradeLevel, FTE, ResidenceStatus, DaysPresent, DaysAbsent, GraduationCohortYear, CohortSize, GraduationStatus, DropoutStatus, LEPEligibility, LEPParticipation, HALEligibility, HALParticipation, SPEDStatus, PrimaryIEPDisability, AlternateAssessment, HighlyMobile, Homeless, Gender, RaceEthnicity, Immigrant, NeSAReadingScore, NeSAReadingProficiency, NeSAMathScore, NeSAMathProficiency, NeSAScienceScore, NeSAScienceProficiency, NeSAWritingScore, NeSAWritingProficiency, CollegeGoing	StudentID, enrollment_begin_date, enrollment_end_date, college_graduation_date, datayears, high_school_grad_date, graduated, college_type_2_4, public_private, degree_title, major, college_name, college_state, college_type_category, agencyid, college_enrollment_status

2.3 Methodology

In this stage of the research, significant attention was given to a proper data sampling process, design, and rigorous data analysis. At the same time, we ensured ethical standards were maintained throughout the study to protect the privacy of students and acknowledged any limitations that may influence the interpretation of our findings.

In order to conduct our analysis, we begun by dividing the original high school data into three high School cohorts based on the years 2012, 2013, and 2014. The term High School cohort refers to the specific year in which students completed their high school education. Similarly, the college dataset was also further subsetted from the original dataset to include only students who attended a 4-year college/university. Students who pursued a combined 2-year and 4-year educational pathway were excluded from the analysis, as this research specifically focused on students who exclusively embarked on a 4-year college journey.

By breaking down the data into these cohorts and using our graduation methodology described above, we were able to determine the expected college graduation years for each cohort groups as 2018, 2019, and 2020 respectively. This breakdown allowed us to narrow down the data and refine it specifically for the research purpose which made it easier to examine the progression and outcomes of high school graduates in Nebraska in terms of their college education. We explored the extent to which high school performance, as measured by factors such as but not limited to school attendance, mobility, Zip Code, ethnicity, and standardized test score as some of the factor to predict college graduation. We also considered the potential influence of demographic factors such as race and gender on this relationship.

In the beginning, we performed a careful manual selection of the variables required from the original dataset for our research by omitting any variables deemed non-contributory to the assessment of high school performance and college graduation outcomes. This careful sub-setting ensured our analysis would be focused and relevant, thereby enhancing the validity of our findings. Specifically, we selected the following variables from each of the dataset.

Handling missing data

Upon completing the variable selection process, we proceeded with a comprehensive analysis of the dataset. We begun by looking at the summary statistics of the data and surprisingly, a substantial portion 4 of the data was found to be missing. Having several missing data can cause the research to be biased, misleading, and also can have challenges in interpretation of the results. So, in order to address missing values and ensure appropriate data integrity, we used a range of techniques for imputation. These methods encompassed Last Observation Carried Forward (LOCF), Mean Imputation, Median Imputation, and Value Imputation, among others. By utilizing these approaches, we aimed to fill in the missing values with the most suitable estimates, thereby enabling comprehensive analyses and preserving the integrity of the data.

Data Preparation

At this point in our research, both of our datasets was in a raw format containing detailed information about students' academic levels in high school and semester level data in college. However, since our end goal is to use Machine Learning algorithms to predict academic success from historical data, we need to consolidate the data to have a single row for each students, capturing their high school and corresponding college data. This consolidation allows for a comprehensive understanding of students' educational journeys and facilitates further analysis in our research.

Several factors have contributed to the occurrence of multiple row values within the dataset. Such as, instances of students switching high schools during the same academic year, grade repetition, and students transitioning between colleges have contributed to the presence of these duplications.

College data preparation

The college dataset comprised a total of 2,226,288 rows, which included various semester-level records such as college name, major, semester enrollment start/end dates, and college type (Private or Public). The dataset specifically pertained to students who pursued higher education after completing high school. To begin our analysis, we initially subsetted the dataset to exclude row data of only those students who were enrolled for a period of approximately two weeks within a given semester. This subset was not selected to account for instances where students may have dropped their courses during a two-week grace period.

Our next objective was to determine whether students exclusively attended a two-year college, four-year college, or both. To achieve this, we introduced a new variable called “college type category” and used the semester level enrollment data to fill in values. This categorization enabled us to focus specifically on students who enrolled in four-year colleges following their high school education, as this constituted the primary focus of our research.

After narrowing down the dataset to meet our specific requirements, our next step involved consolidating the information to obtain a single row representing each student and determining their graduation status. This entailed condensing the multiple rows associated with each student at the semester level into a single row. To accomplish this, we performed a grouping operation based on the student ID and college name. This grouping accounted for situations where a student attended more than one college. Within each group, we extracted the earliest enrollment begin date and the latest enrollment date to capture the duration of their academic journey for the specific college. Furthermore, we incorporated pertinent variables such as the final graduation status, college type (Private or Public), and the corresponding state

In a case where a student went to multiple colleges then we selected the first college graduation status.

As stated in the introductory section of the paper, we proceeded under the assumption that the maximum anticipated timeframe for students enrolled in a four-year college to graduate would be six years (1.5 times the length of the degree or 1.5 times 4). With this in mind, we conducted calculations to determine the total duration of enrollment and introduced a new variable called ‘total graduation years.’ Additionally, we adjusted the graduation status of students who had successfully graduated but took longer than 6 years to graduate to reflect ‘Not Graduated.’ It is important to note that this adjustment was made solely for the purpose of our research analysis.

By following this approach, we obtained a streamlined value of each student’s educational trajectory, taking into account of relevant details.

High School data preparation

To begin with, we have 1,294,297 rows, which contained historical data from high school academic year 2008-2009 until 2019-2020. Our research only focused on students who took NeSA standardized assessments and subsequently went to only 4-year college after High School graduation. To get the desirable subset of the dataset, we performed the following step.

- Created a unique list of student ids who went to 4-year college only from the NSC college dataset.
- We then subset High School dataset to match the student list from the college dataset. This ensured that we work with the data that we need.
- Determined the cohorts in which we have a full end to end data availability

After completing the previous steps, We now have our dataset with 205,737 rows. With the necessary data at hand, our focus now shifts towards the task of consolidating the information into cohesive singlerow records. This approach came with a challenge. The challenge is to determine which row value do we select during consolidation. Our three important dataset variables to perform this process were ‘StudentID’, ‘Datayears’, and ‘GradeLevel’ (see appendix A for variable defination) because we want to initially have a unique row value for a ‘StudentID’ for each ‘Datayears’ and each ‘GradeLevel’. We do not want to have an instance where a students went to the same grade in different academic years or when students move to a different high school on the same academic year. This would create a new record for the same academic year but we only want one record. However, the question is how do we determine what row value do we select? To solve this problem, we introduced a new varable called ‘SchoolDays’ and selected the row value where the student went the maximum. ‘SchoolDays’ is the sum of variables ‘DaysPresent’, and ‘DaysAbsent’ and both of these variables were available in the original High School dataset. We believe that introducing ‘SchoolDays’ and focusing on the school where the student spent the most days can offer us better insights on student’s educational continuity and stability during that academic year.

Now that we have a unique ‘StudentID’ and ‘Datayears, we now proceed to create a unique ’StudentID’, ‘Datayears’, and ‘GradeLevel’. Similarly to what we did in the ‘StudentID’ and ‘Datayears’ by selecting the row value of maximun school days, we do the same for ‘StudentID’, ‘Datayears’, and ‘GradeLevel’. Likewise, we also introduce a new variable called ‘total_repeated_grades’, which counts the total grade a student repeated.

Now that we have established unique identifiers for each student, specifically ‘StudentID’ and ‘Datayears,’ our subsequent task involved expanding this uniqueness to include the ‘GradeLevel’ variable. Similar to the steps we did for ‘StudentID’ and ‘Datayears,’ where we selected the row value associated with the maximum number of school days, we employ a similar approach to determine the appropriate ‘GradeLevel’ for each distinct combination of ‘StudentID’ and ‘Datayears.’ Furthermore, we introduce an additional variable named ‘total_repeated_grades,’ which serves to tally the total number of times a student repeated a grade throughout their academic journey. This could give us an insight if it can be of any influence in college success.

External dataset

To enhance our research, we introduced ZipCode as a variable to our High School dataset. The ZipCode was extraced using the School Name where a student went to High School using Google Places API. Google Places API is service provided by Google Inc.

Overall, We created 6 new variables based on other available variables from High School dataset. New variables such as 'changed_school_same_datayear' which takes the count of number of times the student changed the school in the same academic year and 'total_HighlyMobile_same_datayear' which takes count of number of times student were highly mobile same academic year. the 6 new variables are 'Percent Absent', 'Total Days Present', 'Changed School same Academic Year', 'Total Repeated Grades', 'ZipCode', and 'Highy Mobile same Academic Year'.

Currently, the data for each student is spread across multiple rows, potentially up to a maximum of four rows representing each year of high school. It should be noted that some students may have less than four years due to factors such as transferring from another state or grade acceleration. However, our objective remains to consolidate the student data into a single row.

Removing any of the high school years is not a viable option as it would lead to the loss of significant information, such as gradel level data. To address this challenge, we will proceed with reshaping the dataset by transforming its current long form into a wide form. This transformation will ensure the preservation of all values while achieving the desired structure for our data, allowing us to have a single row representation for each student.

At this point, we had a dataset that was ready to be used for prediction using Machine Learning methods. While it is a big challenge to completely eliminate bias, we did our best to use appropriate techniques to fill in any data that were missing.

2.4 Merging High School Data with College Data

Throughout the previous steps, we conducted a thorough data analysis of a comprehensive dataset, carefully transforming it into a unified dataset that combines the most relevant information in both the High School and college datasets. Additionally, we filtered high school dataset to ensure compatibility with the college dataset, accounting for students who may not have enrolled in college or instances of missing data in the dataset. This step was essential to maintain the integrity and coherence of the dataset, enabling a more accurate analysis of the educational journey from high school to college. Now, our next step is to merge these datasets, effectively encapsulating the entire educational journey from High School to College. This merging process grants us a unique opportunity to investigate the factors that significantly influence college success, leveraging the available data to guide our research. By conducting this analysis, we aim to uncover the key determinants that play a pivotal role in shaping academic achievement during the college years.

We will now proceed to merge these datasets. The College dataset and High School dataset share a common identifier, the student ID. Leveraging this shared attribute, we will merge the two datasets, using the student ID as the key to bring them together. By analyzing the merged dataset, we hope to gain valuable insights into the influential factors that contribute to students' achievements in their college journey.

Following the successful merge, our resulting dataset comprises a total of 12,726 distinct students. These students exclusively pursued a four-year college education and were part of the graduation classes of 2012, 2013, and 2014 from their respective high schools.

In summary:

- 2012 Cohorts have 4,126 students
- 2013 Cohorts have 4,218 students
- 2014 Cohorts have 4,382 students

2.5 EDA / Summary of the Final Data

We are now ready to take a peak at what has been influencing the college success based on the factors we have taken into consideration.

3. Machine Learning Approach

Model selection is an important step in machine learning that profoundly impacts the accuracy and reliability of predictions. Choosing the most suitable model from the array of available algorithms is essential to capture data patterns and generalize well. while selecting the appropriate machine learning model is important, we purposefully aim the research to be more data centric, hence it is also the reason why we have created new variables to support our research. Below we discuss the key consideration involved in the model selection.

For the research, we considered 3 Machine Learning model approach and a Neural Network approach. They are:

Machine Learning

- Random Forest

  Random Forest is a machine learning model that belongs to the ensemble learning family, meaning that it combines multiple individual decision trees models to make predictions. It is robust and effective algorithm for binary classification tasks.

- Gradient Boosting Machines (GBM)

  Gradient Boosting Machines (GBM) is a machine learning algorithm that also belongs to the ensemble learning family, specifically to the boosting technique. GBM combines the predictions of multiple weak models (usually decision trees) to create a strong predictive model.

- Logistic Regression

  Logistic regression is a supervised learning model used for binary classification tasks. It falls under the category of generalized linear models (GLMs) and is widely employed to predict the probability of an event belonging to one of two possible classes.

Neural Network

- Feedforward Neural Networks (FNN)

3.1 Data Prepreparation for Machine Learning

We will begin by ensuring it adheres to the required format for the machine learning model to perform predictions. Currently, the dataset encompasses diverse data types, including numerical, textual, categorical, and temporal. However, certain machine learning models, such as Random Forest, specifically requires the input features (also known as predictive variables) to be in a purely numeric format. This involves representing them as numerical values, such as 0 or 1. Moreover, we will apply this numeric representation to all the models we intend to evaluate for our research. It is worth noting that not all models impose this requirement for numeric representation of the dataset.

The current structure of our dataset looks like this (rows, columns):

- Cohort 2012: (4126, 60)
- Cohort 2013: (4218, 60)
- Cohort 2014: (4382, 60)

Within each cohort, we have a dataset consisting of 60 variables. Out of these variables, 59 are utilized as predictors, while one variable is designated as the target value. The target value specifically captures and denotes the graduation status of the students.

One Hot Encoding

To get the numeric representation of the dataset, we will be using a method called one-hot encoding which converts our dataset into a numerical form.

- What is One Hot Encoding?

  One Hot Encoding is a data transformation technique commonly used in machine learning and data preprocessing to convert categorical data into a numerical format that can be effectively used by machine learning algorithms.

- Why are we using One Hot Encoding on our research dataset?

  The selection of predictors also knows as select variables are mostly categorical. For example: `NeSAReadingProficiency`, `RaceEthnicity`, `Immigrant` are some of the variables that have categorical data. While the data is preserved, it is transformed in a way that is useful to machine learning models.

After applying one-hot encoding to the dataset of each cohort, the resulting dataset exhibits the following shape(rows,columns):

- Cohort 2012: (4126, 425)
- Cohort 2013: (4218, 423)
- Cohort 2014: (4382, 418)

As we can see in the matrix above that the number of columns has expanded following the implementation of One Hot Encoding. This expansion in column count is a direct result of the encoding process, which horizontally expand the variable representation by introducing additional columns. Each discrete or nominal variable lacking inherent order or numerical meaning is allocated a separate column to effectively capture its distinct categories or levels. One important observation to consider is that the column numbers may not be consistent after One Hot Encoding the data is due to the presence of non-numeric data in the dataset. Since these non-numeric values need to be represented in a numeric format, they can result in varying column counts.

In the final phase of data preparation for the Machine Learning model, we will assess the relevance of each variable. In this context, we have decided to drop the 'StudentID' column. While the 'StudentID' was utilized extensively in previous steps to manipulate and refine the dataset, its inclusion in the model is deemed unnecessary since it does not contribute to the prediction process. Consequently, the 'StudentID' column is omitted from the dataset to ensure a more focused and streamlined modeling process.

3.2 Train-Test data split

The train-test data split is performed in machine learning to evaluate the performance of a model on unseen data. By splitting the available dataset into two subsets, namely the training set and the testing set, we can train the model on the training set and then evaluate its performance on the testing set. This allows us to assess how well the model generalizes to new, unseen data and helps us estimate its performance in real-world scenarios. This separation ensures it prevent overfitting, where the model performs well on the training data but fails to generalize to new data. It is a crucial step in model development to ensure reliable and accurate predictions.

For the three distinct cohorts in our dataset, we will be performing a data split using the `train_test_split` method from the `scikit-learn` library. Dataset will be divided 80% for training and 20% for testing the model. To ensure reproducibility, we will set the random state value to 42. This fixed random state value allows us to obtain consistent results and replicate the same data split behavior, which is beneficial for testing, debugging, and comparing different models.

This split dataset will serve as the foundation for training and evaluating all the models. With the completion of this step, we are ready to input the data into various machine learning models and evaluate the performance.

3.3 Machine Learning Report : Random Forest

- What is Random Forest?

  A Random Forest is an ensemble learning technique used for both classification and regression tasks. It is based on decision tree algorithms and builds multiple decision trees during training. Random Forest leverages the power of multiple decision trees to improve predictive accuracy and reduce overfitting. Instead of relying on a single decision tree, Random Forest builds a collection of decision trees and combines their predictions to make more robust and accurate predictions. This ensemble approach helps reduce the variance associated with individual trees and makes the model more resilient and effective in a wide range of tasks, including classification and regression.

  Random Forest provides several useful features that enhance its interpretability and assist in understanding the importance of features in the model. These parameters helps us fine tune the model to extract the best fit considering all variability.

Methodology:

We will be using the full featured dataset which we transformed in earlier steps using One Hot Encoding method. Our data split would be 70% for testing and 30% for training. The next step is to find the best hyperparameters for the model. To achieve this, we will be using GridSearchCV or Grid Search Cross-Validation. It is a technique used in machine learning to find the best combination of hyperparameters for a model. This tool systematically explores the various combinations of these parameters and provides cross-validated results. We used cv=5 within the GridSearchCV, which means that the available data will be divided into 5 equal-sized subsets or folds. The model will be trained and evaluated 5 times, each time using a different combination of 4 folds for training and 1 fold for evaluation. By analyzing these results, we can determine the best parameter configuration that maximizes the model's performance. Subsequently, we can fit the model using the selected parameters, ensuring an optimal configuration for our specific task.

To find the optimal parameter for our model, we have created the following parameters and stored into a variable called `param_grid`

```
param_grid = {
    'n_estimators': [50, 100, 150, 200, 250],
    'max_depth': [3, 7, 10, 12, 15],
    'min_samples_leaf': [5, 10 , 15, 20],
    'min_samples_split': [2, 5, 7, 10, 12],
    'criterion': ['gini', 'entropy'],
    'max_features': ['sqrt', 'log2'],
    'bootstrap': [True],
    'class_weight': ['balanced'],
}
```

where,

- **n_estimators**: The number of trees in the forest. Increasing the number of trees can improve performance but also increases computational complexity.
- **max_depth**: The maximum depth of each decision tree. It controls the depth of the tree and helps prevent overfitting. Setting it to None allows the tree to expand until all leaves are pure or contain minimum samples.

- **min_samples_split**: The minimum number of samples required to split an internal node. It prevents further splitting of nodes with fewer samples, reducing overfitting.
- **min_samples_leaf**: The minimum number of samples required to be at a leaf node. It controls the minimum size of leaf nodes and helps prevent overfitting by limiting the number of samples in a leaf.
- **criterion**: The function used to measure the quality of a split at each node of the decision trees in the ensemble. It helps determine the best feature and threshold to use when splitting a node to create child nodes.
- **max_features**: The number of features to consider when looking for the best split. It controls the randomness in feature selection at each split and helps reduce correlation among trees.
- **bootstrap**: Determines whether bootstrap samples are used to train each tree. Setting it to True enables random sampling with replacement, which is the default and recommended approach.
- **class_weight**: It is used to handle class imbalance, where the distribution of classes in the dataset is uneven.

Each of the above parameters can have **n number** of combination and electing the optimal combination of hyperparameters for Random Forest involves evaluating multiple parameter values. Instead of manually testing each value individually, we utilized GridSearchCV. This powerful tool enables automated hyperparameter optimization by exhaustively searching through the parameter grid. By leveraging GridSearchCV, we streamline the process, save time, and enhance the model's performance by identifying the most favorable parameter configuration based on the chosen evaluation metric.

After conducting GridSearchCV on the datasets of all three cohorts, we obtained the optimal configurations that yielded the best fit for each dataset. The parameters are listed below for each cohorts.

Class of 2012

```
n_estimators=150,
max_depth=15,
min_samples_leaf=5,
min_samples_split= 15,
criterion='gini',
max_features = 'log2',
bootstrap=True,
class_weight='balanced'
```

Class of 2013

```
n_estimators=150,
max_depth=15,
min_samples_leaf=5,
min_samples_split= 15,
criterion='gini',
max_features = 'log2',
bootstrap=True,
class_weight='balanced'
```

Class of 2014

```
n_estimators=150,
max_depth=15,
min_samples_leaf=5,
min_samples_split= 15,
criterion='gini',
max_features = 'log2',
bootstrap=True,
class_weight='balanced'
```

After obtaining the optimal parameters from a range of combinations, the selected parameter configuration is applied to the `RandomForestClassifier()`. The model is then trained by fitting the data to learn the underlying patterns and relationships. Following the training phase, predictions are made on the unseen testing data which we had seperated earlier during the **Train-Test split** phase.

Model Evaluation:

. . . TODOs: after finding good score

For each cohort, the acquired metrics are as follows.

Table 2: Performance Metrics using Random Forest

Performance.Metrics	Cohort.2012	Cohort.2013	Cohort.2014
Accuracy	0.87964	0.89840	0.85840
Precision	0.94699	0.96840	0.94346
Recall	0.92334	0.93840	0.94443
F1-score	0.93501	0.93840	0.91031
AUC-ROC	0.57206	0.61840	0.59711

The features of importance for each of the cohorts are as follows:

Table 3: Top 25 predictor supporting the conclusion

COHORT 2012		COHORT 2013		COHORT 2014	
Performance.Metrics	Score1	Performance.Metrics.1	Score2	Performance.Metrics.2	Score3
Accuracy	0.9384	Accuracy	0.9384	Accuracy	0.9384
Precision	0.9384	Precision	0.9384	Precision	0.9384
Recall	0.9384	Recall	0.9384	Recall	0.9384
F1-score	0.9384	F1-score	0.9384	F1-score	0.9384
AUC-ROC	0.9384	AUC-ROC	0.9384	AUC-ROC	0.9384

Feature Importance Scores:

Random Forest provides a feature importance score for each input feature. These scores indicate how much each feature contributes to the model's predictions. Higher importance scores suggest that the feature is more influential in making decisions within the ensemble of trees.

. . . TODOs: after finding good score

Discussion of Findings:

... TODOs: after finding good score

Conclusion:

... TODOs: after finding good score

Recommendations:

... TODOs: after finding good score

3.4 Machine Learning Report : Logistic Regression

Logistic Regression is a popular and widely used statistical model for binary classification problems. It offers a simpler approach and is used when one or more independent variables (predictors) are available that determine an outcome.

Methodology:

The logistic regression model is defined as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Where:

- p represents the probability of the binary outcome (e.g., 1 for an event occurring, 0 otherwise).
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients associated with the independent variables $X_1, X_2, \ldots, X_k$.
- *log* is the natural logarithm.
- $\frac{p}{1-p}$ is the odds of the event happening (the odds ratio).
- β_0 is the intercept or constant term.

Model Evaluation:

Table 4: Performance Metrics using Logistic Regression

Performance.Metrics	Cohort.2012	Cohort.2013	Cohort.2014
Accuracy	0.87964	0.89840	0.85840
Precision	0.94699	0.96840	0.94346
Recall	0.92334	0.93840	0.94443
F1-score	0.93501	0.93840	0.91031
AUC-ROC	0.57206	0.61840	0.59711

Interpretation of Coefficients:

... TODOs: after finding good score

Discussion of Findings:

... TODOs: after finding good score

Conclusion:

... TODOs: after finding good score

Recommendations:

... TODOs: after finding good score

3.4 Machine Learning Report : LightGBM

LightGBM is another popular method widely used for machine learning problems. It is designed for efficient and high-performance machine learning tasks, particularly in the context of supervised learning for classification problems. LightGBM stands for "Light Gradient Boosting Machine."

The LightGBM model can be mathematically defined as an ensemble of decision trees. Each tree, T_i, is a weak learner that aims to minimize a loss function. The final prediction, $F(x)$, for an input feature vector x, is the sum of predictions from all the trees in the ensemble.

$$F(x) = \sum_{i=1}^{N} T_i(x)$$

Where:

- $F(x)$ represents the final prediction made by the LightGBM model for the input feature vector x.
- $\sum$ indicates summation over all individual trees T_i.
- $T_i(x)$ represents the prediction made by the i-th decision tree in the ensemble for the input feature vector x.

Each decision tree T_i is built sequentially, and the objective is to minimize a loss function, typically defined as a measure of the difference between the predicted values and the actual target values. LightGBM uses gradient-based optimization techniques to determine the best splits at each node of the trees.

The final prediction $F(x)$ is obtained by summing the predictions from all the individual trees. This ensemble approach allows LightGBM to capture complex relationships in the data and make accurate predictions for various machine learning tasks such as regression and classification.

Model Evaluation:

Interpretation of Result:

... TODOs: after finding good score

Table 5: Performance Metrics using LightGBM

Performance.Metrics	Cohort.2012	Cohort.2013	Cohort.2014
Accuracy	0.87964	0.89840	0.85840
Precision	0.94699	0.96840	0.94346
Recall	0.92334	0.93840	0.94443
F1-score	0.93501	0.93840	0.91031
AUC-ROC	0.57206	0.61840	0.59711

Discussion of Findings:

. . . TODOs: after finding good score

Conclusion:

. . . TODOs: after finding good score

Recommendations:

. . . TODOs: after finding good score

Conclusion

. . . . TODOs : need to write the conclusion after good model score