Predicting if a college student will dropout or not is difficult

Loyd Flores

December 28, 2024

1 Introduction

Education is the process of receiving or giving systematic instruction at a school or university (Oxford Languages, 2024). From the inception of higher education in 1088 at the University of Bologna, Italy (Arndt, 2022), to its contemporary forms, education has continually adapted and evolved, keeping pace with the changing landscape of knowledge and discovery. Education remains a cornerstone of societal advancement and personal development, serving as a beacon of enlightenment and empowerment across generations. In the age of information, a college education emerges as an essential foundation for navigating the modern world. According to *The Center for Economic and Policy Research*, blue-collar jobs in the United States have been declining, accounting for only 13.6% of total jobs in 2016 (Baker, 2020), while 75% of new jobs require a degree (Trend, 2022). Furthermore, Bachelor's degree holders earn a median annual US salary of $69,368 and associates degree holders earn on median $50,076, while high school diploma holders only at $42,068 (U.S. Bureau Of Labor Statistics, 2023). In addition to facilitating job prospects and an increase in potential earnings, a college education cultivates personal skills such as critical thinking, problem-solving, time management, and communication; through diverse coursework, individuals refine these abilities, and the collegiate environment fosters enhanced communication skills through activities such as events, group projects, and club participation. Moreover, college offers an expansive network of peers and mentors, allowing for the organic expansion of personal network over the duration of one's academic journey. Despite the widely acknowledged positive correlation between educational attainment and both career prospects and personal development, the United States continues to grapple with an alarming annual college dropout rate of 40% (ThinkImpact, 2021). It is imperative to thoroughly investigate the underlying factors contributing to college dropout rates and discern the motivations behind a student's decision to discontinue their academic pursuits. Moreover, exploring the feasibility of simulating and modeling these factors to make predictive analyses holds significant importance, particularly given the hardships endured by individuals without a college degree. The utilization of such predictive models could be useful in proactively identifying students vulnerable to dropout during their academic tenure. This proactive intervention allows institutions to implement customized support strategies, thereby potentially bolstering retention, graduation, and overall success metrics. Consequently, both educational institutions and students stand to gain from these advancements.

2 Phenomena and Models

Phenomena are observable events that occur in the natural world that capture interest. Examples include the movement of planets, the change in weather patterns, or even the elemental transition of water to ice upon reaching its freezing point. To comprehend such phenomena and make informed predictions, a *model* could be beneficial.

Models serve as simplified representations of reality, providing frameworks for understanding complex phenomena, similar to how a globe serves as a scaled-down version of the Earth, allowing us to study its rotational dynamics on its own axis and to visualize geographical locations of countries and continents. Through the utilization of models, phenomena can be comprehended with greater clarity and simplicity.

The phenomenon being modeled in this paper is the student's decision to persist or drop out of their academic program given the circumstances of their current situation. In this study, we denote the phenomenon's response measurement as y, where y belongs to the set Y representing all possible outcomes, i.e., $y \in Y$. For this particular scenario, y can take only two values: 0 or 1, symbolized as y $\in \{0, 1\}$. Here, **1** signifies that a student has dropped out given their current situation, while **0** indicates the opposite. To measure the response measurement exactly there exists no other possible value for y other than the ones previously mentioned. It is important to note that students are not able to partially drop out. Even if a student chooses to pursue their education on a part-time basis, they are still classified as students who have not dropped out. However, once a student drops out or temporarily takes a leave, the study concludes, and any subsequent actions, such as a student returning to school, are considered beyond the scope of this investigation. The term '*current situation*' refers to a specific point in a student's academic journey and the factors influencing their decision, which can be regarded as the *metrics*, a way to measure things, of this phenomenon. The primary aim of the model is to predict the response based on the prevailing circumstances. As the model seeks to categorize input metrics into two distinct groups $\{0, 1\}$, it falls under the classification of a binary classification model.

3 Student Dropout as a Mathematical model

Functions take inputs and use a specific set of steps to produce outputs. The phenomenon being modeled, mentioned previously, is a function as well. It takes a students' current situation as input then uses a specific set of steps to produce a response $\in \{0, 1\}$. It is a straightforward approach to modeling. At first glance, this approach seems direct, offering a clear path from input to output. However, upon closer examination, it becomes obvious that this simplicity causes a fundamental problem. The inputs, which represent a student's current situation, are not properly quantified. Even if situations from millions of students were available, the output may not always be meaningful. Imagine having a pantry filled with unlabeled ingredients: while you have a wide variety of items, without knowing what each ingredient is or how they should be combined, you're unable to create a meaningful dish. Without accurate quantification of inputs, the model's output may fail to capture the intricacies of real-world scenarios, potentially resulting in unreliable or misleading outcomes that cannot be verified. The absence of methods to verify the validity of the response presents a significant challenge, undermining the credibility and usefulness of the model in accurately representing the phenomenon it intends to model.

Luckily, there exists a solution to this problem. *Mathematical models* describe systems by establishing relationships between inputs and outputs. They serve as a method to represent and explain real-world systems and phenomena through a mathematical and empirical approach. Since the model operates within the framework of mathematics, it is imperative that the input data is accurately quantified. The primary challenge lies in effectively quantifying the diverse range of student situations. Without precise quantification, the model may struggle to interpret and process the input data accurately. Quantification entails assigning numerical values or metrics to various aspects of the student's situation, including academic performance, socioeconomic background, and personal circumstances. Just as the response phenomenon is constrained to $\{0, 1\}$, the input data must also adhere to mathematical principles. This ensures that the model is unambiguous and operates within a well-defined mathematical realm, facilitating accurate analysis and interpretation. Once an empirical approach is deployed, namely a mathematical model, the phenomenon being studied could be assumed to be deterministic i.e.

$$y = t(z_1, z_2, \dots , z_q)$$

The equation above implies that the decision of a student to drop out (y) is influenced by several causal drivers ($z_1, z_2, ..., z_q$). These causal drivers are combined using a mathematical function or methodology represented by t. Specifying t as a mathematical function is essential because we are operating within the domain of mathematical models. This ensures that our analysis remains grounded in rigorous mathematical principles, allowing for precise modeling and interpretation. Lastly, it facilitates the portability of the model to other universities, enabling its reuse and applicability in diverse educational contexts. To illustrate, identifying reasonable causal drivers is crucial for predicting whether a student will drop out or not because they provide insights into the underlying reasons behind such outcomes.

Some examples of these causal drivers include:

z_1 = Mental health of student

z_2 = Peer influence on academic performance

z_3 = Total assistance provided by the family

z_4 = Burnout rate

z_5 = Hours of effective study

z_6 = The level of student enthusiasm or passion for their chosen major

z_7 = Difficulty of current classes z_8 = Level Financial of Security

After defining the casual drivers as quantifiable metrics, the groundwork is laid for the emergence of a mathematical model to represent these metrics effectively. However another problem occurs, the casual drivers are not always available. For instance, z_1, which represents the mental health of the student, poses its own challenges. Mental health is multifaceted and can be measured in various ways, making it inherently complex. Moreover, not all students may be willing to participate in psychological assessments.

Additionally, both z_2 and z_3 introduce complexity due to the involvement of other individuals. The influence of peers (z_2) on academic performance is often oversimplified; while it's commonly believed that spending time with studious peers leads to better academic outcomes, reality is far more nuanced. Although surrounding oneself with diligent individuals can foster a culture of learning, various other factors come into play, and it's erroneous to assume a universal correlation.

On the other hand, z_3, which pertains to the extent of assistance provided by the family, underscores the intricate composition of familial support. Monetary aid is undoubtedly significant, but the implicit forms of support are equally invaluable. Actions such as providing food, engaging in check-up conversations, and offering mental encouragement contribute significantly to a student's well-being and academic performance. The complexity of family dynamics extends beyond financial contributions, highlighting the multifaceted nature of familial support in a student's educational journey.

Moreover, z_4, z_5, z_6, z_7 (Burnout rate, Hours of effective study, Level of enthusiasm, and Difficulty of current classes) are subjective and unique to each student. The difficulty of a class can vary significantly from one student to another due to a multitude of factors. Consequently, students may allocate different amounts of time to study effectively, introducing another challenge: how to quantify the effectiveness of study time and burnout rate. Because of subjectivity, there is no uniform way of quantifying mentioned metrics in such a way that it is normalized across all students.

The limitations of the present day may also impede the quantification of casual drivers. However, as humans progress, they may discover methods to quantify and normalize the variables (z's) associated with these drivers, potentially leading to improvements in understanding. It's important to note that the phenomenon of interest is not stationary, meaning its statistical properties change over time. This lack of stationarity complicates modeling efforts, as past models may not accurately predict future responses. As highlighted in the introduction, educational institutions have continuously evolved over time. The challenges faced in education vary as time progresses, and additional factors may also contribute to enhancing the student experience. Because the causal factors (z) being overly complex and are not stationary, the rules governing their combination (t) become even more intricate. The solution to this is by approximating t and coming up with a simpler function called f. Further, the casual drivers also have to be approximated. Expressed mathematically, $x_1, x_2, ..., x_p$ will act as proxies for the casual driver $z_1, z_2, ..., z_q$. In light of these approximations, we introduce a new equation:

$$y = t(z_1, z_2, \ldots, z_q) \approx f(x_1, x_2, \ldots, x_p) + \delta \text{ , where } \delta = \text{(t - f)}$$

Based on the context provided earlier in this section, the new equation represents a model for predicting student dropout (y). This prediction relies on a multitude of factors denoted as casual drivers (z's), which collectively influence the outcome. However, these factors are intricate and not entirely quantifiable, expressed through the function $t(z_1, z_2, ..., z_q)$. Given the complexity and elusive nature of these factors, our understanding is limited, leading us to approximate their influence with another set of variables x’s, represented by the function $f(x_1, x_2, ..., x_p)$. The $x's$ are called independent variables or simply features. These variables play a crucial role in the predictive model, as they are the quantifiable attributes or characteristics that are believed to have an impact on the outcome variable. The symbol ’$\approx$’ is used to signify this approximation, indicating that our model may not capture all nuances accurately.The discrepancy between the actual outcome (t) and our approximation (f) is represented by δ, denoting *the error due to our incomplete understanding or ignorance*. This error, $\delta = (t - f)$, highlights the unaccounted aspects that contribute to the difference between the predicted and actual outcomes.

Therefore, while our model provides insights into student dropout tendencies, it’s essential to recognize its limitations and the inherent uncertainty stemming from the approximation process. As aptly stated by *George Box*, ’All models are wrong, but some are useful.’ Models are inherently approximations, and their utility lies in their ability to explain phenomena better as they approximate closer to reality. Since, the casual drivers are too complex they are approximated using features. Below are the features identified as effective in capturing the diverse characteristics and behaviors of casual drivers along with their data type:

x_1 = Total credits obtained by the student (Numeric)

x_2 = College Major (Nominal)

x_3 = Current GPA (Numeric)

x_4 = Current Credits being taken (Numeric)

x_5 = Total cost of semester in USD (Numeric)

x_6 = Amount of Federal Financial Aid received in the current semester in USD (Numeric)

x_7 = Amount of Financial Support for the current semester given by the family in USD (Numeric)

x_8 = Amount of Scholarships received for the current semester in USD (Numeric)

x_9 = Total student loan debt in USD (Numeric)

x_{10} = Highest educational attainment in the family (Ordinal with 7 levels)

x_{11} = Commute time in hours (Numeric)

x_{12} = Hours dedicated to external obligations (Numeric)

x_{13} = Student income per semester in USD (Numeric)

Each feature is now readily accessible in mathematical format, aligning with our model's predefined mathematical structure. It's evident that the chosen features may significantly influence a student's decision to drop out during their current semester. The numerical values all represent continuous variables, such as distance in miles and amounts in USD. Categorical features are features that can be categorized into two subgroups: ordinal and nominal. Nominal features lack a specific order, whereas ordinal features exhibit a ranking system within their classes, within this study only nominal features are used.

The following features, x_1, x_2, x_3, x_4, offer insights into students' academic performance, providing an understanding of their current context or situation. Depending on their chosen major and the number of current credits taken, we gain insight into what students are experiencing. For example, a student enrolled in a full load of 15 credits but pursuing a relatively undemanding major (subject to individual interpretation) may find their situation less challenging compared to someone taking 9 credits in a more rigorous major. However, this isn't always the case. Factors such as current GPA and total credits earned also play a crucial role. Students with higher GPAs or those who are nearing the completion of their academic programs, having accumulated more credits, typically demonstrate a stronger aptitude for managing academic responsibilities effectively, reflecting their readiness to handle heavier workloads and demonstrating greater accountability. These metrics collectively provide valuable insights into students' academic performance and their ability to manage various academic demands.

In addressing the economics of education, x_5, x_6, x_7, x_8, and x_9 play pivotal roles as determinants influencing a student's likelihood of dropping out. Financial constraints often serve as compelling factors driving such decisions, especially as tuition fees continue to escalate. It doesn't help that the United States is considered to be the most expensive country for students in 2023 (Staff, 2023). For instance, for the 2023-2024 academic year, private universities experienced a staggering 4% increase in fees (Poolos, 2024), while public four-year colleges saw a 3.0% rise (Nietzel, 2023). The total cost of the semester (x_5) provides a clear indication of the economic burden students face, while x_6, x_7, x_8, and x_9 explain how students personally navigate this financial strain. A combination of these features can significantly clarify a student's decision to drop out due to financial constraints. For instance, students benefiting from scholarships and robust financial aid packages, effectively rendering their education costs minimal or even providing financial surplus, are less likely to abandon their studies. Conversely, those lacking assistance or accumulating substantial debt may find it impractical to continue or have no viable alternative but to discontinue their education.

Measuring the highest educational attainment within a family (x_{10}) is important, as it serves as a significant factor in preventing student dropout rates. Parents with higher educational achievements tend to instill values that prioritize academic success in their children. Conversely, individuals with minimal degrees or no education may lack the resources or support systems necessary to encourage academic perseverance. For example, if both parents have PHDs, the likelihood of a student dropping out could be assumed to be low, given the conducive educational environment typically associated with highly educated families (Nelson, 2019). In this study, this feature encompasses seven possible levels, ranging from 0 to 6, with 6 representing the highest educational attainment and 0 indicating the lowest. These levels are as follows: PHD and above (6), Masters (5), Bachelors (4), Associates (3), Other, which encompasses a variety of credentials such as Trade School, Boot camps, etc. (2), High School (1), and None (0).

To estimate burnout percentages, x_{10}, x_{11} and x_{12} are utilized. These features offer insights into external factors that may contribute to burnout. For instance, if students reside far from the school and have a low monthly income, this could potentially lead to dropout scenarios. Such students may be engrossed in work commitments due to their low income, and the distance from school could exacerbate the challenge of attending classes regularly. Furthermore, the number of external obligations like extra activities can impact dropout probabilities. Engaging in numerous non-academic obligations, such as parenting responsibilities, touring performances, or professional fighting, may overwhelm students and detract from their academic focus. Conversely, participating in school-based extracurriculars like sports teams or club activities can foster a sense of belonging and routine, potentially reducing the likelihood of dropping out by integrating school into students' daily lives.

All the aforementioned features are readily accessible through the registrar's office in most universities/colleges. However, it's worth noting that some features, such as x_7, x_8, x_9, x_{10}, x_{11}, and x_{13}, may not always be available and may require direct input from the student. These features are succinct and accurate representations of various aspects of student situations. With a total of 13 features, denoted by p, it's generally advisable for the number of rows, or total student situations, represented by n, to be larger than p. While a higher n is desirable, a value too close to p may still lead to unusable data, as it may fail to capture the variance of our dataset adequately. This phenomenon, known as estimation error, will be further discussed in subsequent sections.

4 Learning from Data

Now that the features and their collection methods are defined, the next step is to streamline the process of understanding them. To tackle this challenge, an empirical technique called supervised learning, utilizing historical data should be deployed. *Supervised learning* involves extracting patterns from past data, specifically the provided features, by employing algorithms to map inputs $(x_1, x_2, ..., x_n)$ to outputs (y). In this process, the mathematical model previously defined transforms into a supervised machine learning model called g. This transformation occurs as the model is processed by a computer, utilizing both the defined inputs and the designated algorithm, defined as

$\mathcal{A}$. This model, g, is trained to discern patterns from past data, enabling it to make predictions on current situations, even on data it hasn't been explicitly trained on or *unseen data*. This is beneficial for understanding the phenomenon of interest due to the intricate nature of the issue. Once the model, denoted as g, is developed and implies learning, it can be particularly useful when early signs are detected in a student. The model g could be employed to predict potential dropouts by collecting essential features from the student, and inputting the features into g, allowing intervention to occur before the undesirable outcome materializes.

4.1 Training Data D

It was previously noted that casual drivers (denoted as z) were approximated by features (x) for utilization in mathematical models. Additionally, it was emphasized that employing an empirical approach via supervised learning models is the optimal method. To accomplish this, it's imperative to appropriately structure the features for the computer to construct a supervised learning model (g). This process of organizing the data results in the formation of the *training dataset*, referred to as **D**. This dataset comprises all feature columns $(x_1, x_2, ..., x_{13})$, collectively labeled as X, representing the entirety of features. These features are then linked to a single output, denoted as y. In other words, each row of inputs in X corresponds to one entry in y, which simply means, a student's situation leads to a decision. For instance, if there are 30 students in the dataset, there will be 30 corresponding combinations of X and y. Mathematically, represented as $D =< X, y >$. This structuring enables the algorithm to interpret the feature values and establish a correlation with the provided output. To paint a clearer picture here is an example of 1 entry:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	y
60	Pure Math	2.9	15	10000	0	0	0	25000	1	20	10	500	1

This specific scenario could be interpreted as follows: The mentioned student is a sophomore with 60 credits, pursuing a major in Pure Mathematics, with a GPA of 2.9, just barely maintaining their academic standing. They face a significant financial burden, with each semester costing $10,000 and no financial support from government, family, or scholarships. Additionally, the mounting student loan adds to their financial stress, and living far from campus exacerbates their challenges. Despite working 10-hours a week and earning a monthly salary of only $500, the student struggles to cover expenses, ultimately leading to their decision to drop out, as indicated by the outcome variable $y = 1$.

The larger the dataset, with more students represented by rows of X's and y's, denoted as n, the better the performance of the function g will be. Consider the scenario described above as the sole training data for the model, where $n = 1$. In such a case, the model might wrongly generalize that all students facing similar situations will drop out, which may not always hold true. Conversely, not all students in favorable circumstances necessarily graduate. When a model fails to generalize accurately, this error is termed as *estimation error*, which could be mitigated by gathering more data.

5 Support Vector Machines (SVM)

As discussed previously, the algorithm $\mathcal{A}$ serves as a tool for machines to interpret data. Support Vector Machines, also known as Maximum Margin Classifiers, are algorithms designed to discern patterns in the variable y by identifying a clear boundary between data points. This binary classification approach mirrors situations like the legal age limit for purchasing alcohol in the United States, which is a strict yes/no scenario. Consequently, SVMs are well-suited theoretically to predict outcomes such as whether a student will drop out of school, given their ability to handle binary decision-making effectively. Since $p = 13$ the threshold will be harder to visualize. To visualize it is necessary to use illustrations with a smaller size of p.

Here is what SVM will look like if $p = 1$:

The information extracted from Figure 1 originates from the Iris dataset, containing details on various species along with their characteristics. In this context, the primary determinant for identifying a particular iris species is the sepal length. The analysis reveals a clear pattern: irises with a sepal length ranging from 0cm to 5.2cm are predominantly of the Setosa species, while those

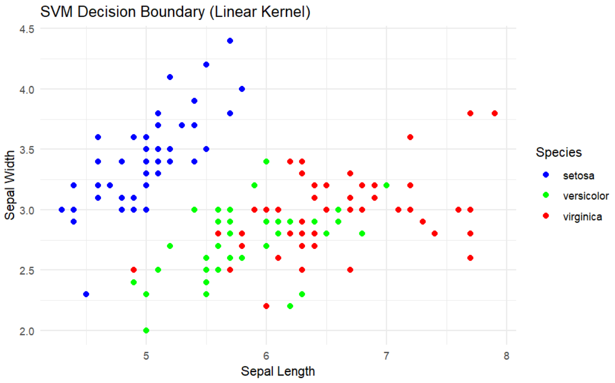


Figure 1: Threshold for sepal length on Iris dataset

measuring between 5.3cm and 5.9cm tend to be Versicolor. Instances surpassing 5.9cm typically correspond to the Virginica species. This observation underscores the effectiveness of support vector machines (SVM) in discerning these distinctions.

Now that the idea of splitting the data using a threshold is clear, the next step is to visualize with more features and understand how SVM really shines, here is an example where p = 2:

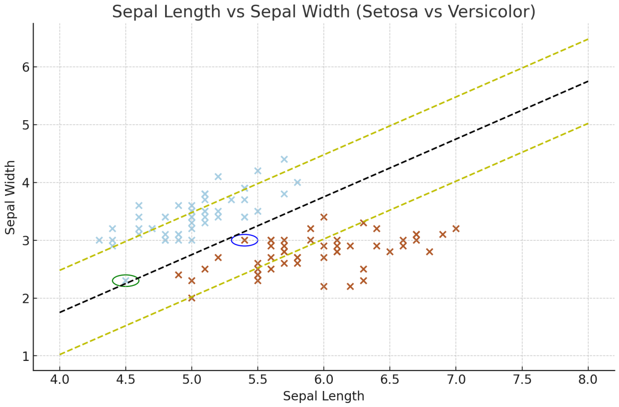


Figure 2: Threshold for sepal length and sepal width

In this example, the focus is on visualizing the context closely related to the phenomenon under study, wherein the model is configured to predict from two classes, specifically, understanding the Iris classes: Setosa and Versicolor. SVM achieves this by utilizing support vectors, as depicted in Figure 2. Support vectors assist in adjusting the position of the Maximum Margin Hyperplane, which is the central line, by constraining the area it can be placed in. As the objective is to separate the two classes, the support vector machine offers a theoretical guideline, referred to here as l_{upper}

and l_{lower}, indicating where the maximum margin hyperplane can be positioned. It's termed as the maximum margin hyperplane because the aim of the support vector is to maximize the space it provides to the line. This space, termed the "wedge," implies the separation of the two classes, with the line being positioned where it optimally fits and achieves the most accurate separation or minimal error.

Returning to the scenario of this issue with $p = 14$, visualizing it can be challenging due to the high dimensional involved. Instead of appearing as a line, it would manifest as a plane in a three-dimensional space. Now that there is a clear visual picture of how Support Vector Machine works, the formula to produce model g could be defined as the following:

$$y = t(z_1, z_2, \ldots, z_q) \approx f(x_1, x_2, \ldots, x_p) + \delta \approx \text{g, where } g = \mathcal{A}\ (\mathbf{D}, \mathcal{H})$$

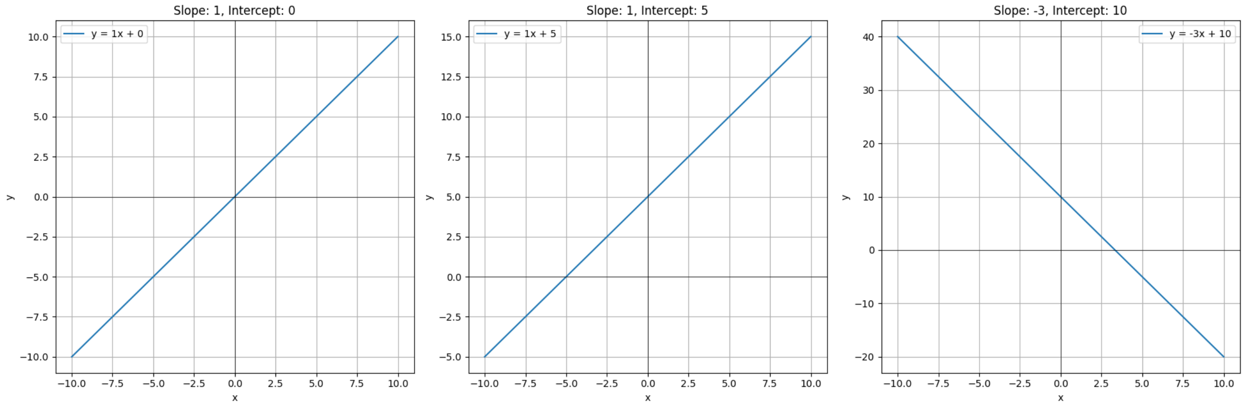
In this equation, the output y of a phenomenon is represented as an approximation of a function f acting on variables $x_1, x_2, \ldots, x_p$, with an added term δ. To enable computational processing, the variables are transformed into a suitable format represented as a dataset **D**. An algorithm $\mathcal{A}$, such as a Support Vector Machine, learns from this dataset to generate a model g, as f might be too complex or inaccessible. However, the choice of potential models $\mathcal{H}$, known as the candidate set, could approximate it and is yet to be defined.

5.1 Candidate Set $\mathcal{H}$ for SVM

Various algorithms operate differently, necessitating a distinct set of candidate functions, denoted as $\mathcal{H}$. In the case of Support Vector Machines (SVM), this set comprises functions aiming to maximize the margin, or wedge which is the space in between the two different classes. SVM then places a line within the wedge; hence, $\mathcal{H}$ represents the set of all possible hyperplane (decision boundaries) that separate the data points in a given feature space.

Defining the candidate set would look like this: $\mathcal{H} = \{\ 1\ \vec{w} \cdot \vec{x}\ \text{- b} \geq 0 : \vec{w} \in \mathbf{R}^p, \text{b} \in \mathbf{R}\ \}$

The condition 1 $\vec{w} \cdot \vec{x} - b \geq 0$: defines the decision rule for the SVM. It states that a data point $\vec{x}$ belongs to one class if the dot product of the weight vector $\vec{w}$ and the feature vector $\vec{x}$, minus the bias term b, is greater than or equal to 0. $\vec{w}$ and the intercept b allow the line to move towards the best fit, the movement of a line could be shown using a set of graphs with different slope and intercept values :


Slope: 1, Intercept: 0

Slope: 1, Intercept: 5

Slope: -3, Intercept: 10

Figure 3: Movement of the line with different slope and Intercept values

Weight Vector ($\vec{w}$): Represents the coefficients, or slopes in Figure 3, of the hyperplane in feature space. Learned via optimization techniques like *gradient descent* during training to maximize margin between classes while correctly classifying data. Each feature contributes differently to the decision boundary, so we use a vector to represent the weights assigned to each feature. This vector allows us to capture the combined influence of all features on the decision boundary in a compact and efficient manner.

Bias Term (b): Allows shifting of hyperplane away from origin. Learned alongside weight vector during training. Usually denoted as b or b_0, distinct from class labels.

$\mathbf{R}^p$: p-dimensional real number space, where p is number of features.

$\mathbf{R}$: Real number line.

5.2 Sources of Error

Given that the model (g) endeavors to approximate f by identifying h^*, a function within $\mathcal{H}$ that minimizes error, any additional approximation introduces further error. The precise definition of g to equate to y is as follows:

$$y = g(x) + e,$$

where e encompasses both ϵ (estimation error) and δ (ignorance). To write y in full it would look like this:

$$y = g(x) + \underbrace{(g(x) - h^*(x))}_{\text{\#3 estimation}} + \underbrace{(f(x) - h^*(x))}_{\text{\#2 misspecification}} + \underbrace{(t(x) - f(x))}_{\text{\#1 ignorance}}$$

The two types of errors have been previously defined. The first type, ignorance, encompasses factors that cannot be accounted for due to the intricate nature of the world. The second type, estimation error, arises when the model fails to capture the underlying patterns of the data, a situation that might be mitigated by collecting more data. The error that remains undefined is misspecification, which occurs due to the complexity of the data and the limitations of the selected function h by the algorithm $\mathcal{A}$. As previously mentioned, h^* represents the function with the least error, closest to f, but achieving this ideal may not always be feasible for the model. For example, if the data exhibits nonlinear patterns, then a hypothesis space ($\mathcal{H}$) that includes curved functions (h) becomes necessary. The model would incur considerable misspecification error if $\mathcal{H}$ consists solely of linear functions.

Before tackling the issue, it's important to acknowledge that the primary source of error could fluctuate depending on the circumstances. For instance, in cases of low amounts of data, the predominant error might stem from estimation. Conversely, when dealing with non-linearly separable data, the most significant error could arise from misspecification. However, if the dataset is ample and linearly separable, yet the Support Vector Machine (SVM) still yields a considerable error, it likely reflects ignorance regarding the problem, necessitating more refined approximations ($x's$).

5.3 Error Metrics for SVM

In the previous subsection, we discussed the types of errors that can occur in a binary classification problem. However, it's crucial to understand how these errors are measured. Various metrics can be used to evaluate the performance of a binary classification model. These metrics help guide the modeling process by indicating areas for improvement and adjustment within the model. It is gauged by comparing the actual value of the phenomenon y to the model's prediction $\hat{y}$.

The metrics to be used in this study include:

Accuracy: The proportion of correctly classified instances out of the total instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision: The proportion of true positive predictions out of all positive predictions, emphasizing the correctness of positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Sensitivity): The proportion of true positive predictions out of all actual positive instances, highlighting the model's ability to capture positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1 Score: The harmonic mean of precision and recall, providing a balanced measure of model performance considering both precision and recall.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In summary:

High Accuracy: Good overall performance in correctly classifying instances.

High Precision: Minimizing false positives, important when the cost of false positives is high.

High Recall: Minimizing false negatives, important when the cost of false negatives is high.

High F1 Score: Balanced performance in terms of precision and recall, suitable when you want to balance between false positives and false negatives.

5.4 Vapnik Objective function

If the data is not linearly separable, SVM may perform poorly and incur significant misclassification errors. Fortunately, a solution exists. The Vapnik objective function, denoted as λ, is introduced as a hyper parameter, which can be adjusted. This function determines the degree to which outliers should be penalized. When dealing with only a few outlier data points, incorporating the Vapnik objective function allows for the best possible fitting of a line, accommodating outliers without requiring excessive adjustments to the line itself. With the Vapnik hyper parameter added the the SVM's equation for the model becomes :

$$g = \mathcal{A}(\mathcal{D}, \mathcal{H}, \lambda)$$

6 K Nearest Neighbors (KNN)

What if the data lacks linear separability and contains numerous outliers? In such cases, the fundamental principle of Support Vector Machines (SVM) is undermined. As previously mentioned, there are various algorithms available, with K Nearest Neighbors (KNN) being one of them. KNN operates differently by employing a distance function to group data points together, using the hyper parameter k. For instance, when $k = 1$, it groups points individually, whereas with $k = 3$, it considers the three closest neighbors for each point. A visual aid may better illustrate how k affects the grouping process:

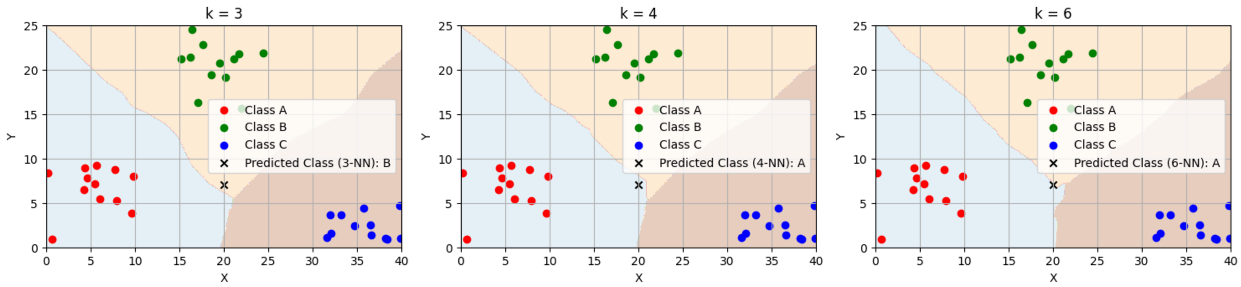


Figure 4: KNN with different values of k

Based on Figure 4 when k = 3 the new point x is classified into class B, yet when k = 4 and k = 6, it is classified into class A. As the hyper parameter is changed the way the points are grouped together will change.

6.1 Distance Function

The main premise of KNN is finding the k nearest neighbors but how does it actually figure out which points reside the closest. There are more many ways to define such distance function such as Manhattan distance but in this study Euclidean distance is preferred as it is the default for this algorithm. It works like this :

$$\text{Euclidean distance} = \sum_{j=1}^{p}(x_{i,j} - x_{*,j})^2$$

p represents the number of dimensions or features in the space.

$x_{i,j}$ refers to the j-th component of the i-th point.

$x_{*,j}$ refers to the j-th component of another point (often a reference or centroid).

This formula calculates the sum of the squared differences between corresponding components of two points in p-dimensional space. By summing these squared differences and taking the square root, you obtain the Euclidean distance between the two points.

Since KNN utilizes a completely different algorithm and does not require a set of candidate functions, the model utilizing KNN as an algorithm will be defined as :

$$g = \mathcal{A}(\mathcal{D})$$

6.2 Sources of Error and Error Metrics for KNN

Given that the problem remains within the realm of binary classification, we will employ the same evaluation metrics utilized in SVM for assessing KNN: Precision, Accuracy, Recall, and F1 score. Delving further into analysis, one notable challenge that KNN might encounter, apart from inherent ignorance, is estimation error. This occurs when insufficient data impedes KNN's ability to discern the genuine pattern during training, possibly resulting in poor performance when applied in real-world scenarios, particularly if the value of k is set too low due to limited available data.

7 Overfitting and Underfitting

Given that all aspects, from the training data and algorithm for data comprehension to the resultant model produced by machine learning, as well as the method for assessing its performance, are clearly outlined, it appears that the trajectory is straightforward. However, there remains a crucial concern: the metrics previously established may be inaccurate, potentially leading to a misleading perception of exceptional performance.

Returning to the training dataset $\mathcal{D}$, it serves as the basis for training the model g. However, a significant issue arises when the model initially achieves a low score; it becomes tempting to continuously retrain it until it achieves flawless error metrics. Yet, these metrics may not accurately reflect real-world performance.

Consider this scenario: picture a student preparing for an upcoming test by solely focusing on a practice exam. This student devotes an extensive amount of time to studying the practice exam, to the point where they could answer its questions with ease, even with their eyes closed. However, on the actual day of the exam, they discover that the problems presented are all modified variations of those in the practice exam. Consequently, despite their thorough preparation, the student performs poorly. Rather than grasping the fundamental concepts underlying the material in the practice exam, the student hyper fixated on specific problems, leaving them ill-prepared for unseen variations of the questions.

Before delving further, it's important to note that the metrics previously established are all in-sample, meaning they are assessed solely within the confines of the data that has been seen before. In the context of the example provided, this pertains to the practice test. On the other hand, out-of-sample evaluation gauges how the model performs in real-world scenarios, where it encounters unseen examples. In the example's context, this corresponds to the actual exam, where the student faces new challenges beyond what was previously encountered in practice.

The previously provided example illustrates over fitting, a phenomenon where a model becomes overly focused on the nuances of the training data, leading to the false impression that it will perform well. This is reflected in high in-sample metrics. However, when exposed to unseen data, the model performs poorly, resulting in low out-of-sample metrics.

Conversely, under fitting can be likened to studying only half of a practice exam, wherein you struggle to perform adequately even within that limited scope. Consequently, when faced with the entirety of the actual exam, performance is expected to be significantly worse. Underfitting is characterized by obtaining a notably low in-sample score, indicating that the model fails to grasp the underlying patterns within the data, resulting in similarly poor out-of-sample performance.

8 Model Validation

The approach to addressing this issue is commonly known as *train-test split*, where the dataset $\mathcal{D}$ is divided into two distinct subsets: D_{train} and D_{test}, with $\mathcal{D} = D_{train} + D_{test}$. This method facilitates model training on D_{train} to assess its performance using in-sample metrics, providing an estimate of how the model might generalize to unseen data. Subsequently, D_{test} serves as a set of unseen data, accessed only once at the conclusion of modeling, providing an unbiased evaluation of the model's true performance. Typically, D_{train} comprises a random 80% to 90% of the dataset, while D_{test} accounts for the remaining 10% to 20%.

A solid starting point involves comparing the model to g_0, which serves as the null model, providing a baseline estimate. If the machine learning-generated model performs worse than g_0, it indicates a significant error in the process, whether it stems from feature selection, approximating the z values in Section 3, or choosing the appropriate algorithm for the task. Given that the desired output y belongs to the set 0, 1, the null model is defined as follows:

$$g_0 = \text{mode[y]}$$

which means that predictions will simply reflect the most frequent class. For instance, if the dataset predominantly comprises dropouts, the null model will predict that every new entry is a dropout.

9 Model Selection

Considering the multitude of potential model variations achievable through minor adjustments, such as altering the λ value in SVM or the k hyper parameter in KNN, each adjustment yielding vastly different models, how can one ascertain that the model generated by the machine learning pipeline is indeed the optimal one? This dilemma is commonly referred to as the ”model selection problem.” Since all conceivable models are essentially approximations, theoretically imperfect, yet some might prove practical, and furthermore, the randomness inherent in the train-test split may yield sub optimal or less useful models. To solve this a technique called *K-fold-cross validation* is deployed.

Now that the process of splitting $\mathcal{D}$ into three subsets: D_{train}, $D_{validation}$, and D_{test} is clear, it is essential to know that k-fold cross-validation complements this strategy. Instead of relying solely on a single split, k-fold cross-validation offers a systematic approach to iteratively cycle through different partitions of D_{train} and $D_{validation}$.

In k-fold cross-validation, we further divide D_{train} into k folds. During each iteration, the model is trained on k-1 folds and validated on the remaining fold. It's important to note that all the models trained on different algorithms and hyper parameters are trained on different fold combinations. This ensures that each data point is utilized for validation exactly once, enhancing the reliability of the evaluation process.

Throughout the k-fold cross-validation iterations, the errors or performance metrics of all these models are recorded. Once the k-fold cross-validation process is complete, the metrics are then aggregated to obtain a comprehensive assessment of the model's performance. By leveraging all available data for training and validation in a systematic manner, k-fold cross-validation provides honest and stable evaluations, thereby enhancing the overall reliability of the model's performance estimation. In addition to assessing model performance through techniques like k-fold cross-validation, it's crucial to consider how well our models generalize beyond the observed data points. This consideration brings us to the concepts of extrapolation and interpolation.

Extrapolation involves making predictions beyond the observed data range, which can lead to unreliable forecasts due to reliance on assumptions about unobserved data points.

Interpolation is the estimation of values within the range of observed data, providing insights into how well models generalize within known data boundaries.

When evaluating models using techniques like k-fold cross-validation, considering both extrapolation and interpolation helps gauge their ability to make accurate predictions within and beyond the observed data range. In the end, the model selected will be the combination with the lowest aggregated errors.

10 Making Predictions

Throughout each step of the machine learning pipeline, fundamental problems arose, yet solutions were fortunately at hand. Now, at the culmination of this pipeline, the imperative is to devise g_{final}. This model represents the optimal combination generated within the pipeline, yielding the lowest out-of-sample errors. Should this combination happen to be SVM, it will feature the best λ; otherwise, if it were KNN, it would possess the best k. This ideal combination is then fed the entirety of $\mathcal{D}$ to maximize the data points it could learn from, thereby minimizing estimation error. Once this process concludes, g_{final} stands ready for use and deployment, enabling predictions with the lowest possible error attainable with the present knowledge and data available at the time of training.

11 Conclusion

The quote **”All models are flawed, but some may prove beneficial”** warrants repetition here. Given the inherent limitations of the process at hand, it becomes clear that all efforts made thus far are mere approximations of reality. Compounded by the intricate and ever-changing and non-stationary nature of predicting student dropout behavior, even a highly effective model today might falter tomorrow. Achieving accurate predictions necessitates ongoing model maintenance, continual data gathering, and deeper exploration of the underlying causal factors behind dropout decisions. After investing significant time and effort, only to achieve fleeting accuracy, it becomes evident that predicting student dropout outcomes is an inherently challenging task.

References

Arndt, G. (2022, January 2). The history of academic degrees. *Everything Everywhere*. Retrieved from https://everything-everywhere.com/the-history-of-academic-degrees/

Baker, D. (2020, January 30). The decline of blue-collar jobs, in graphs. *Center for Economic and Policy Research*.
Retrieved from https://cepr.net/the-decline-of-blue-collar-jobs-in-graphs/

Nelson, J. K. (2019). *IMPACT OF PARENT EDUCATION ON STUDENT SUCCESS*. Utah Valley University, Orem.

Nietzel, M. (2023, November 2). Average College Tuition Increased Less Than Inflation For 2023-24. *Forbes*.
Retrieved from: https://www.forbes.com/sites/michaeltnietzel/2023/11/02/average-college-tuition-increased-less-than-inflation-for-2023-24/?sh=4e0a9446496f

Poolos, O. (2024, February 8). University announces tuition increase for 2024-2025, largest in last decade. *Student Life*.
Retrieved from: https://www.studlife.com/news/2024/02/01/university-announces-tuition-increase-for-2024-2025-highest-in-last-decade

Staff, S. I. (2023, May 17). *10 most expensive countries in the world for students*. Retrieved from: https://studyinternational.com/news/most-expensive-countries-in-the-world/

Trend, D. (2022, August 8). 75% of new jobs require a degree while only 40% of potential applicants have one. Truthout.
https://truthout.org/articles/75-of-new-jobs-require-a-degree-while-only-40-of-potential-applicants-have-one/

ThinkImpact. (2021). College Dropout Rates. Retrieved from https://www.thinkimpact.com/college-dropout-rates/

U.S. Bureau Of Labor Statistics. (2023, September 6). Employment Projections. *U.S. Bureau of Labor Statistics*. Retrieved from https://www.bls.gov/emp/chart-unemployment-earnings-education.htm