**Deliverable 3: Classification of High Default Risk in U.S. Higher Education Programs**

By: Brian Hunt, Logan Laszewski, Eliza Shane

***Introduction***

In the ever-evolving landscape of higher education financing, student loan default rates have become a critical point of concern for policymakers, educational institutions, and students. With the rising cost of education, many students increasingly rely on federal loans to fund their academic pursuits. In the context of federal student aid, a borrower is typically considered to have defaulted if no payment has been made for 270 days, triggering significant financial consequences. Recent proposals have emerged suggesting that borrowers who fall behind on their student loan repayments could face a mandatory wage garnishment—up to 15% of their disposable income could be deducted directly from their paychecks. This approach is intended to enhance collections on defaulted loans by automating a portion of the repayment process (Forbes, 2025). While this measure will alleviate some of the burden on lenders, it raises the stakes for incoming college students if enacted. Addressing this issue not only informs better loan disbursement practices but also aids students in making financially sound educational choices.

Understanding the factors that most strongly influence default risk is imperative not only for reducing personal financial hardships but also for maintaining institutional compliance with federal regulations. This project aims to address this challenge by using data mining techniques to identify educational programs and institutional characteristics most strongly associated with default risk. By analyzing data from the U.S. Department of Education’s College Scorecard, we develop predictive models to discern patterns linked to loan repayment challenges. The insights from this project can support educational institutions in improving financial aid strategies and guide students in selecting programs less likely to lead to financial hardship and make informed post–secondary education decisions.

The central questions guiding this research are:

1. *Which types of educational programs are most strongly associated with future loan default?*
2. *What institutional factors can serve as early indicators of default risk?*

***Business Understanding***

The problem of student loan defaults presents a multifaceted challenge for both educational institutions and students. When students default on their loans, the consequences extend beyond individual financial hardship, impacting the institutions that granted the degrees. Schools with high default rates face federal penalties, including loss of Pell Grant Eligibility and other forms of financial aid. This not only affects their financial health but also their ability to attract new students. As a result, there is a pressing need for institutions to better understand the risk factors associated with default and develop proactive strategies to mitigate them.

Student loan defaults are influenced by a complex interplay of factors, including program type, credential level, region, economic conditions, institutional characteristics, and the nature of educational experience (e.g., remote versus in–person). Despite the extensive data collected on student loans, stakeholders often struggle to interpret these variables in a way that translates into actionable strategies.

Private lenders care about this information because they understand the link between academic pathways and repayment behavior. They do not always or often ask, but in the context of certain specialized loan agreements, like income share agreements, the lenders do often inquire about the field of study, along with the other standard questions.

The federal government is concerned with these numbers because of the substantial amount of loans it funds through the Direct Loan Program. College Scorecard gives them the ability to hold institutions accountable for providing a fair value proposition between the student (the consumer) and the university.

The reason that the federal government and universities care is intertwined, though the university's underlying motivation may be less immediately obvious. Universities risk losing eligibility for federal aid if their default rates remain too high, which directly impacts enrollment and institutional revenue.

Data mining offers a powerful solution by enabling the extraction of patterns from large, multifaceted data sets. Unlike traditional statistical methods, machine learning algorithms can account for complex interdependencies and reveal subtle correlations between educational characteristics and loan outcomes. The ability to not only predict default risk but also explain the underlying reasons behind these predictions is essential for both risk mitigation and decision support.

Ultimately, the goal of this project is to create a strong analytical framework that enables stakeholders to understand and address the predictors of student loan default. By grounding our recommendations in empirical evidence, we aimed to promote financial stability for individual borrowers and long-term sustainability of educational institutions.

***Data Understanding and Preparation***

To address the problem of student loan default effectively, we utilize comprehensive data from the U.S Department of Education College Scorecard, specifically focusing on the ’Most Recent Data by Field of Study’ and ‘Most Recent Institution-Level Data‘ datasets. These datasets are publicly available through data.gov, providing both accessibility and credibility to our analysis. The most recent cohort's field of study data set contains detailed program-level information, while the Most Recent by Institution dataset offers contextual data on the institutional environment.

Data preparation involved merging these data sets using UNITID as a common key, allowing us to link program-specific data with broader institutional attributes. The integration was essential to contextualizing program outcomes within as many institutional environments as possible, enhancing the models.

We systematically handled missing data and privacy-suppressed entries (marked as ‘PS’) by removing rows containing these entries. To simplify analysis, we grouped academic majors into broader categories, which proved to be one of the more difficult portions of preprocessing. We consolidated these into ten distinct CIPDESC\_Category groupings and then created dummy variables from those. This transformation was needed, as our main goal was to explore the relationship between the field of study and loan default risk. Using all unique majors would have created hundreds of dummy variables, making the model unwieldy and reducing interpretability. Grouping majors also allowed clearer patterns to emerge.

We also applied one-hot encoding to categorical variables, such as credential level (CREDDESC), institution type (CONTROL)—where we mapped Public, Private for-profit, and Private non-profit to Public and Private—and geographic region (REGION). The outcome variable (BBRR3\_FED\_COMP\_DFLT) was transformed into a binary classification: programs weigh a default rate above 10% versus those below. This involved transforming the outcome variable, BBRR3\_FED\_COMP\_DFLT (Percentage of undergraduate completer undergraduate federal student loan borrowers in default after 3 years), which originally appeared as a mix of ranges and inequalities (e.g., "<=0.2", "0.11–0.14", "<=0.10", etc.). With the help of Dr. Su Dong, we standardized this variable into numeric bins that could be used for classification modeling.

Several related variables were explored but not used as predictors in the final model, as they reflect outcomes only observable well after the time of loan origination. These include DEBT\_ALL\_STGP\_ANY\_MDN (median Stafford and Grad PLUS loan debt), EARN\_MDN\_HI\_1YR (median earnings one year after graduation), and EARN\_PELL\_WNE\_MDN\_1YR (earnings of Pell Grant recipients), among others. While these would be useful for exploratory analysis and context, they were excluded from our model-based project due to their limited relevance in early-stage default prediction.

Additionally, we had a data imbalance (significantly more students *not* defaulting than defaulting) that created imbalanced results when we experimented with true penalties like 15% and 30%. As a result of our 10% cutoff, class distribution was fairly balanced, about 52% below 10% and 48% above, which helped avoid common issues associated with similarly imbalanced classification problems.

By addressing data quality and integration challenges systematically, we established a powerful foundation for building predictive models, allowing us to explore the relationship between institutional character characteristics and student loan default risk.

***Modeling***

To accurately predict the probability of student loan default risk, we employed data mining techniques that balance predictive accuracy with interpretability. Given the complexity of the problem—identifying which educational programs and institutional characteristics most significantly influence default rates—we needed models that could capture nuanced relationships while also providing clear insights. After careful consideration, we selected Logistic Regression (with Lasso and Ridge regularization) and Random Forest as primary modeling approaches.

***Modeling Selection and Rationale***

Logistic Regression was chosen for its ability to provide clear interpretability through coefficients and p-values, allowing us to quantify the impact of each predictor on the likelihood of default. Logistic Regression is particularly well-suited for binary classification tasks, where the objective is to model the probability of an event occurring (in this case, loan default). We used Lasso (L1) and Ridge (L2) Regularization to address multicollinearity and reduce overfitting. Lasso helped with feature selection by shrinking less significant coefficients to zero, while Ridge minimized the impact of correlated variables, leading to more stable model estimates.

While Logistic Regression offers high interpretability, its assumption of linearity between predictors and the log odds of the outcome can be limiting when dealing with non-linear data patterns. To overcome this limitation, we also implemented Random Forest, an ensemble learning method that constructs multiple decision trees and aggregates their predictions. Random Forest is particularly effective for capturing complex, non-linear interactions within data, making it complementary to the more interpretable Logistic Regression model. Moreover, Random Forest provides feature importance metrics, allowing us to identify which variables most strongly influence default risk.

***Model Implementation***

Our primary outcome variable, BBRR3\_FED\_COMP\_DFLT (the percentage of undergraduate federal student loan borrowers in default after 3 years), was transformed into a binary classification variable, indicating whether the default rate exceeded 10%. This threshold was chosen because it aligns with the warning territory of federal guidelines that categorize programs exceeding this rate as high risk.

Logistic Regression was implemented with both Lasso and Ridge regularization to improve the model stability and enhance feature selection. The model was trained using cross-validation to assess consistency across different data subsets. We configured Random Forest with multiple trees to enhance generalizability and hyperparameters optimized to balance model complexity and computational efficiency. By using both models, we could assess both linear and nonlinear relationships, providing a comprehensive analysis of default risk factors.

***Why These Models Address the Business Problem***

The combination of Logistic Regression and Random Forest directly addresses the business problem by offering both predictive power and interpretive clarity. Logistic Regression helps educational institutions understand the specific variables that most significantly influence default risk, enabling targeted interventions and policy adjustments. For instance, identifying that certain credential levels or program types correlate with higher default rates can prompt schools to reassess their curricula or offer additional financial counseling.

Random Forest complements this by capturing complex non-linear relationships that may be missed by linear models. Its ability to handle variable interactions makes it ideal for analyzing combinations of factors like regional characteristics and program delivery methods that jointly affect default outcomes.

By integrating both models, we leverage their respective strengths—interpretability from multimethod Logistic Regression, and pattern detection from Random Forest—providing a strong foundation for truly understanding and addressing student default risk.

***Evaluation***

Evaluating the performance of our predictive models is crucial to determining their practical utility and ensuring that their insights translate effectively into real-world applications. Given the high stakes associated with student loan defaults, including financial and reputational consequences for both individuals and institutions, it was imperative to select evaluation metrics that accurately reflect model performance while balancing trade-offs between precision and recall.

***Evaluation Metrics***

We prioritized precision and recall as our primary metrics, given their relevance to our model and problem. Precision measures the proportion of high-risk programs out of all programs predicted to be high-risk. This metric is particularly significant for educational institutions because falsely labeling a program as risky can result in unwarranted changes to the curriculum offered or the financial aid applied. High precision ensures that when our model flags a program as high-risk, it is accurate and trustworthy.

Recall, in contrast, captures the proportion of actual high-risk programs that the model successfully identifies. This metric is crucial because failing to detect genuinely high-risk programs can leave students vulnerable to financial distress.

Balancing precision and recall is inherently challenging because improving one often reduces the other. In this context, we used F1-score as a secondary metric to assess how well the model balances these two aspects. While precision and recall are individually valuable, the F1-score harmonizes them, providing a single value that reflects the model’s ability to handle class and balance. However, our primary focus remained on interpreting precision and recall independently, as they directly addressed institutional and student needs.

***Model Evaluation and Comparison***

Logistic Regression and Random Forest were evaluated using precision, recall, and F1-score, allowing us to compare their effectiveness in identifying high risk programs. Logistic Regression demonstrated higher precision, which makes it suitable when false positives need to be minimized. This model’s interpretability also enabled us to directly analyze the impact of specific variables, offering insights that could be directly translated into policy changes or program adjustments.

In contrast, Random Forest demonstrated higher recall, making it more effective when the primary objective is to capture as many high-risk cases as possible. The model’s ability to detect complex interactions between variables, made it particularly valuable in identifying nuanced patterns that may not be immediately apparent. Although it is less interpretable than logistic regression, its ability in detecting defaults proved beneficial in identifying broader risk trends.

***Other Methodologies Attempted and Considered***

As mentioned, our project focused on building and comparing several modeling approaches, including Logistic Regression and Random Forest classifiers. We also spent time experimenting with neural networks, which produced slightly stronger predictive results. However, given our project’s emphasis on interpretability and understanding feature contributions, we chose not to move forward with neural networks due to their "black box" nature.

At one point, we considered modeling the default rate as a continuous outcome variable. However, since the BBRR3\_FED\_COMP\_DFLT column included a mix of ranges, inequalities, and suppressed values, this introduced inconsistencies and limited the feasibility of regression-based approaches. This led to our decision to convert the variable into categorical bins and focus on classification instead—an approach more compatible with both the data format and our goal of identifying high-risk program profiles.

We also considered using more advanced ensemble methods such as XGBoost. While potentially more powerful, we chose not to pursue these methods given our project’s focus and the scope of tools covered in class. Additionally, Random Forest already provides the level of flexibility and performance needed to meet our objectives. We also wanted to avoid overfitting and maintain transparency in our analysis, especially given our goal of providing insights that schools and students could act on.

Ultimately, the combination of Logistic Regression and Random Forest enabled us to both predict student loan default risk and explain the factors driving it. This supports our business objective by offering actionable, data-backed guidance for institutions and individuals.

Although collection agencies can add fees of up to 25% on defaulted federal loans (LegalClarity, 2025), the aggregate financial impact of defaults remains negative for the U.S. Treasury, most private lenders, and certainly for borrowers, who actively are trying to contribute to the economy through study. Our model frames ROI from the perspective of institutions and students, who face reputational harm, declining enrollment, and reduced lifetime earnings when defaults rise. This remains true even if a small segment of the collection industry temporarily profits from default-related fees.

***Deployment***

To support deployment, findings should be communicated in a clear, interpretable format using visual tools and summary reports that highlight key risk factors such as region, credential level, and major category. For example, our appendix includes visualizations showing how default risk varies across degree types (Visual 2), regions (Visual 3), and fields of study (Visual 4). These can be shared through presentations, reports, or dashboards to make insights digestible for both technical and non-technical audiences.

We recommend developing an interactive dashboard that allows users to input key attributes, such as program type or region, and view estimated default risks based on model results. This tool would empower students to assess financial risk and help institutions track high-risk programs efficiently. It would also simplify the use of College Scorecard data, which is otherwise large and difficult to navigate. Being able to use a dashboard to tailor a college search to a personal loan situation in the privacy of your own home, or while on a college visit at the financing office, would create a more ideal situation for all incoming students and institutions.

These insights can support multiple stakeholders. Students can use them to weigh program choices against financial risk, while institutions can apply them to advising, support services, and resource planning. Private lenders may incorporate program-level risk into underwriting to improve fairness over relying solely on school-wide default rates. Policymakers could also use this data to flag high-risk programs early and encourage schools to act before hitting critical CDR thresholds.

However, there are several challenges and considerations related to deployment. One key issue is the need for responsible interpretation. Because our data is at the program level, not the individual level, the insights should not be used to make assumptions about individual borrowers. Misapplication of the results could lead to unfair generalizations or biased decision-making.

There are also important ethical considerations. Some variables, such as region or credential type, may correlate with systemic inequities. Without proper context, using these results to steer students away from certain programs could inadvertently reinforce existing inequalities. To address this, any deployment should include disclaimers, documentation, and—ideally—collaboration with advising professionals who can contextualize the results in conversations with students. Additionally, there is a privacy risk for small programs, as previously mentioned. A way to work around the privacy suppression to paint a fuller picture while simultaneously maintaining privacy is a serious hurdle.

In terms of risk mitigation, we recommend that the findings be used as a starting point, not a definitive answer. Users should supplement model outputs with local knowledge, institutional data, and student input before making decisions. The models should also be updated periodically as more recent College Scorecard data becomes available to ensure continued relevance and accuracy.

Finally, while our current work focuses on broad institutional trends, it also highlights the need for deeper research into why default risk varies across programs. Deployment should go hand-in-hand with follow-up efforts to investigate root causes and develop targeted support strategies—not just flagging problems but working toward solutions.

***Return-on-Investment Perspective (Student Centric, Institutionally Mission Driven)***

Creating a dashboard would go beyond compliance and venture into full transactional transparency. When institutions publicly highlight programs with strong student ROI—and actively improve or sunset those that lag—they send a clear market signal of accountability. Over time, that transparency attracts more applicants, strengthens alumni satisfaction, and bolsters external rankings, all of which translate into higher enrollment yield, stronger donor support, and a more resilient business model. This reinforces the dual benefit: supporting students' financial outcomes while strengthening the institution’s long-term positioning.

***Conclusion***

The issue of student loan default is complex, involving varied responsibilities and financial implications for students, institutions, and lenders alike. This intricacy underscores the need for clear, actionable insights derived from reliable data. By carefully processing and analyzing the messy and nuanced data from College Scorecard, our project has laid the foundation for future predictive models and dashboards that can inform all stakeholders. We aim to support data-driven decision-making that ultimately enhances educational, financial, and macroeconomic outcomes.

**Appendix**

U.S. Department of Education, Federal Student Aid. “Default.” *Federal Student Aid*. https://studentaid.gov/manage-loans/default. Accessed May 5, 2025.

Forbes article, "Education Department Will Resume Student Loan Debt Collection In May" (April 21, 2025).

[Education Department Will Resume Student Loan Debt Collection In May](https://www.forbes.com/sites/shaharziv/2025/04/21/student-loan-borrowers-beware-wage-garnishment-resumes-in-may/)

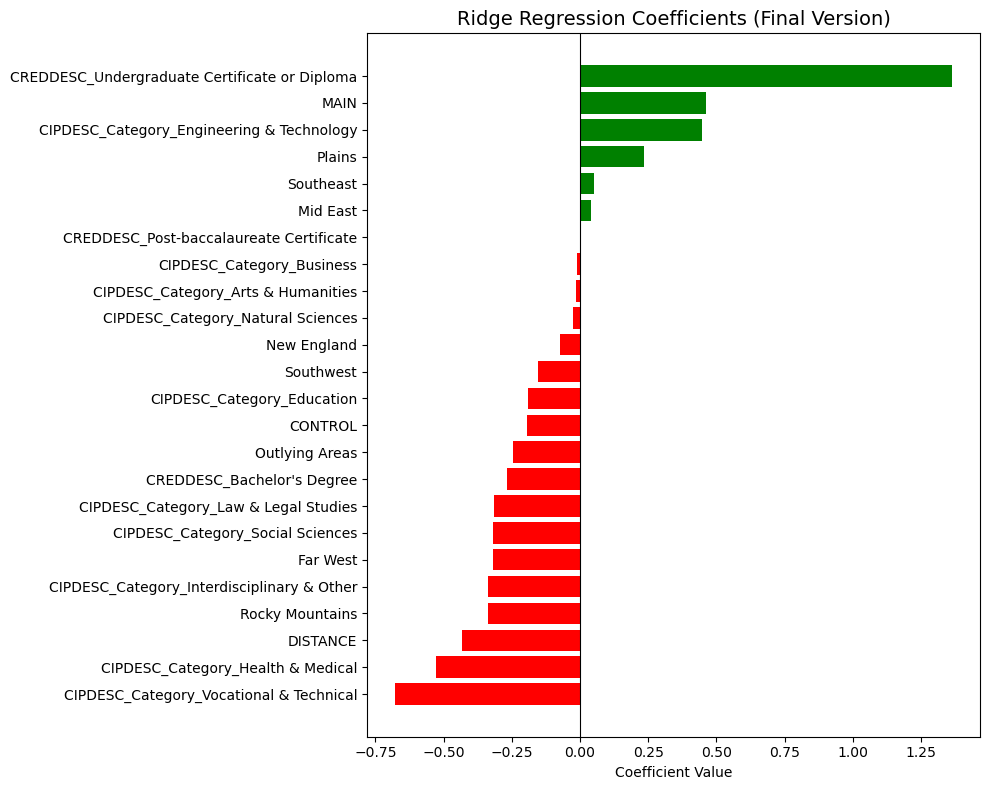
Fredman, J.. (2019, November 19). *Policymakers should change how student loan defaults are measured*. National Association of Student Financial Aid Administrators (NASFAA).<https://www.nasfaa.org/news-item/20100/TICAS_Policymakers_Should_Change_How_Student_Loan_Defaults_Are_Measured>

Federal Reserve Bank of Cleveland. (May 1, 2008) *Trouble Ahead for Student Loans?*

<https://www.clevelandfed.org/publications/economic-commentary/2008/ec-20080501-trouble-ahead-for-student-loans>

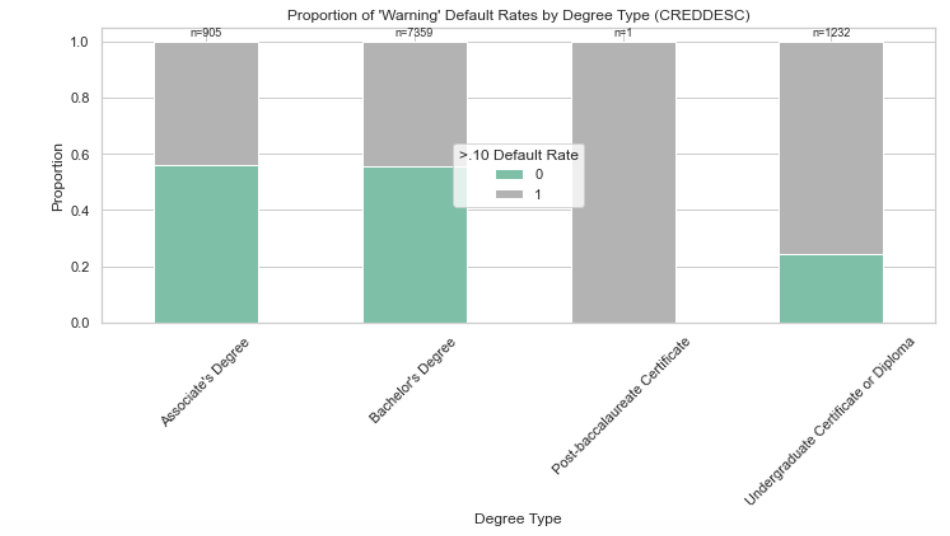
LegalClarity. (2025, February 13). *Can a debt collector charge more than the original debt?* LegalClarity.<https://legalclarity.org/can-a-debt-collector-charge-more-than-the-original-debt/>

**Visual 1**

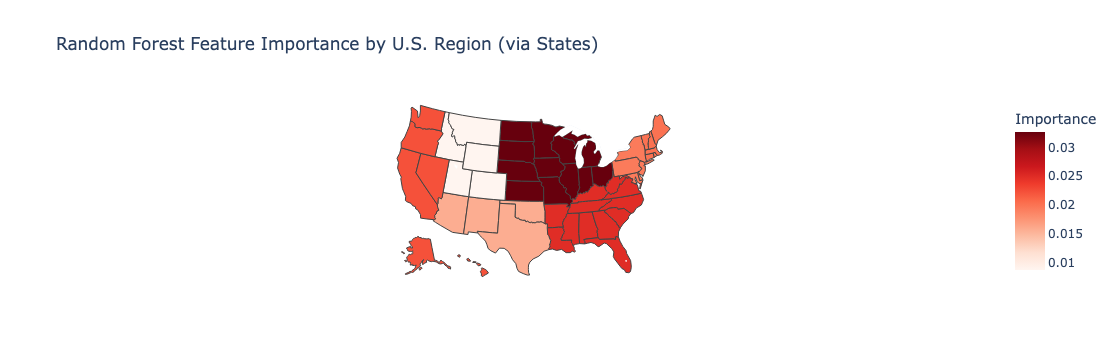


*This bar chart shows how default rates vary based on the type of academic program, the credential received, and region, using the Ridge Regression Coefficients. One takeaway is that students receiving undergraduate certificates or diplomas are at significantly higher risk of defaulting than students receiving a Bachelor’s Degree.*

**Visual 2:**

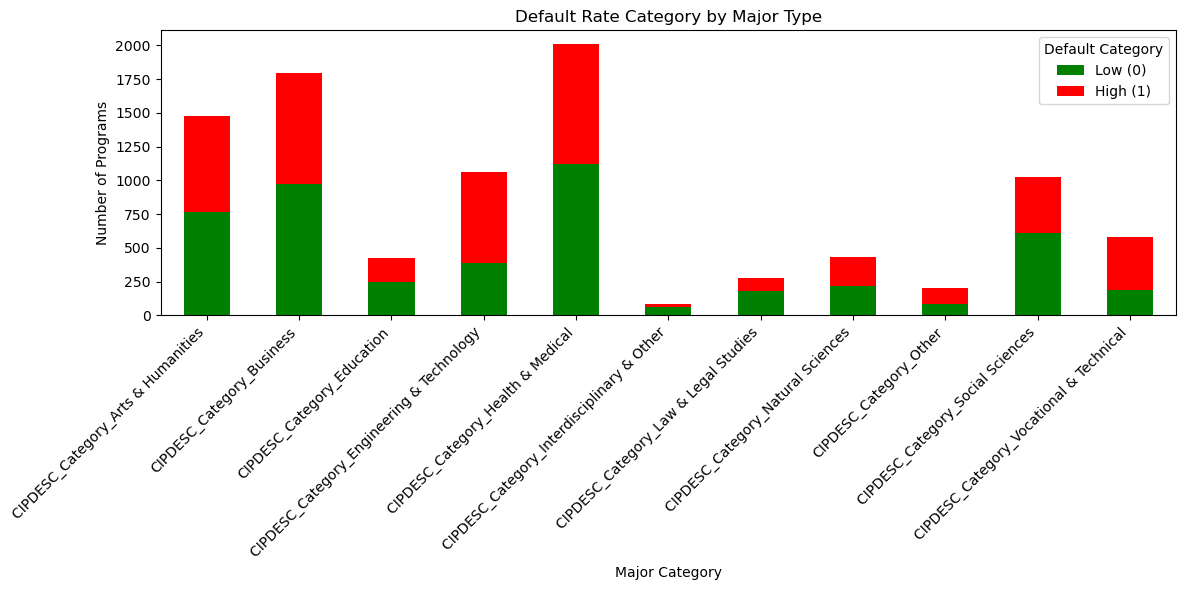


*This bar chart highlights the high-risk level of an Undergraduate certificate or diploma, followed by Associate’s degrees. Bachelor’s degrees have the lowest proportion of default rates above 10%, reflecting the stability of repaying these loans compared to the other degree types. Overall, this visual supports our conclusion that degree type is a major predictor of federal loan default risk.*

**Visual 3: **

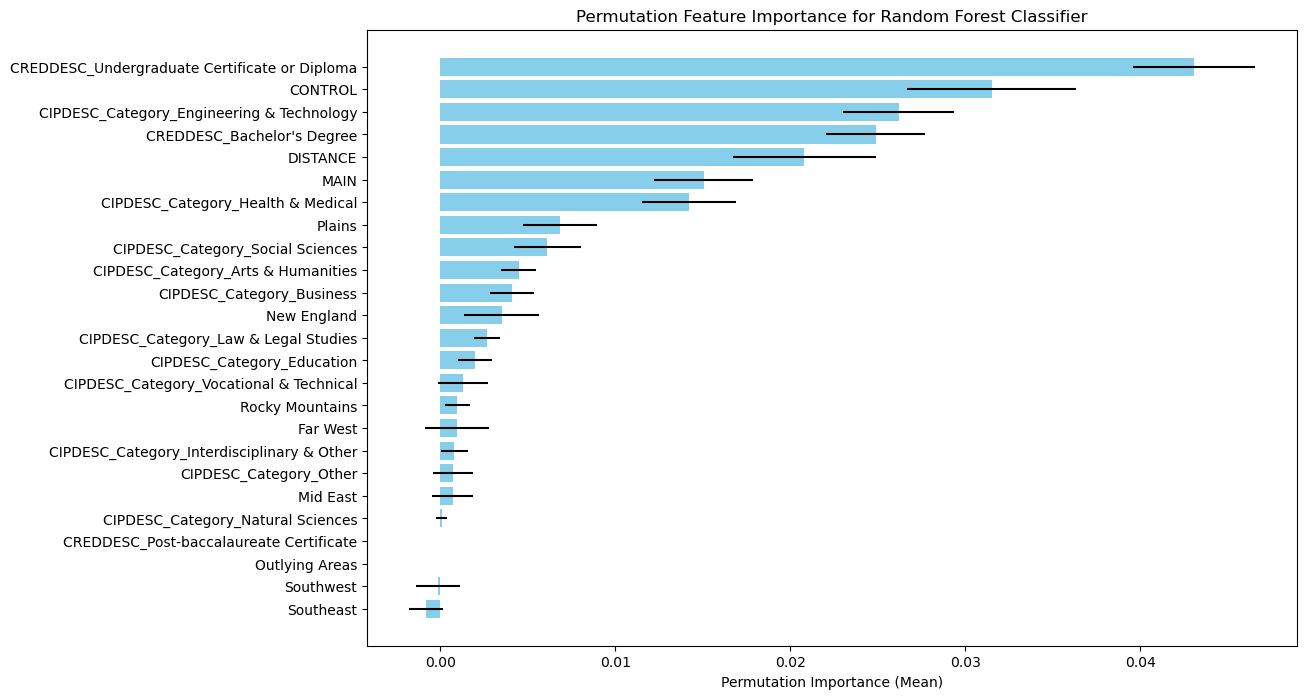
*This regional map, using the Region variable from the merged Institution dataset, highlights the Random Forest feature importance for each region in the United States. This visual demonstrates how each region correlates with the percentage of programs at a higher risk for defaulting on federal loans. The map shows that the Great Lakes region and the Plains region have more programs above the 10% default threshold than programs in other regions of the U.S., such as the Southwest.*

**Visual 4:**

****

*This visual shows the proportion of programs within each major category that exceed the high-risk defaulting threshold of 10%. Fields such as Engineering & Technology and Arts & Humanities show a larger proportion of high-risk programs compared to fields of study such as Social Sciences and Education. These insights should be taken into account by prospective students in fields with higher default risks.*

**Visual 5:**



*This visual shows the permutation importance of various predictive features in the Random Forest model used to classify high default risk in college programs. Permutation importance measures how much a model’s accuracy decreases when a specific feature’s values are randomly shuffled, meaning it shows how critical that feature is for making accurate predictions.*

*The length of the thin black line indicates the variability in the importance score across multiple permutations. A shorter line means the feature’s importance is more consistent, while a longer line indicates greater variability. If a feature has a high importance but a wide confidence interval, it suggests the model’s reliance on that feature might not be stable. Conversely, a narrow confidence interval indicates the feature is consistently contributing to accurate prediction.*

*Features such as credential level (e.g. Undergraduate Certificate or Diploma) and institution type (public or private, labelled as CONTROL) exhibit the highest importance, indicating their strong influence on predicting default rates. Additionally, fields like Engineering & Technology, and Health & Medical fields also show significant contribution to model performance.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Brian – Business Framing, Bridging Code, and Final Writing

Brian took the lead in defining the business problem, ultimately guiding the group toward focusing on student loan default classification by high default risk. He explored potential datasets and identified the College Scorecard as a viable source for data mining. Early on, he wrote the initial code framework that set the foundation for the rest of the analysis, enabling Logan to build on it and expand the modeling. Brian also collaborated with Eliza on organizing the presentation slides and later played a key role in editing and refining the final paper. Brian also added the permutation visualization as well as the precision and recall comparisons of all models. Brian also added and annotated Visual 5.

Logan – Lead Coder and Data Visualizer

After Brian established the initial coding structure, Logan dove deeper into understanding the data and developing the predictive models. He made additions and changes to the code, created visuals for the analysis, and met virtually with Professor Dong to resolve questions around handling the outcome variable and model type (categorical vs. numerical). While primarily focused on the technical implementation, Logan also contributed bullet points and notes that served as a starting point for the paper and presentation.

Eliza – Primary Writer and Presentation Designer

Eliza served as the group’s lead writer and slide designer. She worked with Brian to build the slide deck and developed detailed speaker notes to support the presentation. For the final paper, she transformed group notes and bullet points into well-written paragraphs, ensuring consistency and limiting repetition. In addition, she created the map visualization in Python to support our analysis.