**IEOR4524 Final Report**

**Chilean Highway Toll Collection**

**Sponsor**: Ardian Private Equity

**Group 32:**

Michelle (Meixi) Peng ([mp4270@columbia.edu](mailto:mp4270@columbia.edu))

Kris (Tiantian) Chen (t[c3315@columbia.edu](mailto:c3315@columbia.edu))

Vidhi Agrawal ([va2504@columbia.edu](mailto:va2504@columbia.edu))

Howard (Haozhou) Yu ([hy2818@columbia.edu](mailto:hy2818@columbia.edu))

Raphael Barthes ([rpb2141@columbia.edu](mailto:rpb2141@columbia.edu))

1. **Introduction**

**1.1. About the Sponsor**

The sponsor of this project is a leading global private equity firm known for its investments across various sectors, with a focus on infrastructure. The firm is dedicated to leveraging technological advancements to enhance performances of its infrastructure investments. This commitment reflects a strategic emphasis on integrating data-driven solutions and innovative practices to optimize operational efficiencies and ensure sustainable growth within its portfolio. The firm's proactive approach in adopting cutting-edge technologies underlines its vision to set new benchmarks in infrastructure management and investment returns.

**1.2 Problem Background**

This project is centered on addressing the operational challenges in free-flow tolling systems. Specifically, the project addresses issues of toll evasion or "leakage," which have been exacerbated following the global pandemic, increasing from 4% to 20%.

Previous efforts to address these challenges have primarily involved the use of a basic regression model that focused on individual characteristics to predict default rates. This approach was characterized by a narrow scope of data extraction and a subjective selection of variables, limiting its effectiveness. Importantly, the results from these initial models were not integrated into decision-making processes for enhancing toll collection strategies, rendering them insufficient in contributing to the operational improvements needed. The need for a sophisticated, data-driven approach to enhance the financial sustainability of toll operations is pressing, given the strategic importance of these assets in the sponsor's investment portfolio.

**1.3 Project Goal**

The overarching goal of this project is to refine urban toll management through the application of sophisticated data analytics. There are two primary objectives: firstly, to develop a predictive model using toll transaction data. This model will forecast payment defaults and enhance the accuracy of payment timing estimates, thereby optimizing the involvement of collection agencies and improving the efficiency of toll revenue collection. Secondly, the project aims to perform an analysis of how macroeconomic trends impact payment delays. By linking economic factors with payment behaviors, this analysis will provide insights to help stakeholders anticipate changes in defaults, and inform their negotiations regarding toll pricing with the Chilean government.

Success for the predictive model will be quantitatively measured by its accuracy in forecasting payment defaults and the precision of payment timing predictions, which in turn will enhance revenue collection efficiency. For the macroeconomic analysis, success will be determined by the clarity and strength of the correlations identified between economic variables and payment delays, which will play a critical role in strategic negotiations for toll rate adjustments. Key deliverables of the project include a fully operational predictive model, a comprehensive macroeconomic correlation report, a 2-slide summary of our work, and 2 detailed presentations for the sponsor and Vespucio, respectively, that explain the findings and strategies of both the predictive and correlation analyses. Expected outcomes from these deliverables include improved toll collection efficiency, as well as robust, evidence-based economic rationale for overall default anticipation and toll rate renegotiation.

1. **Data and Resources**

## **2.1. Data Collection Methodology**

Data collection utilized two channels. The primary source is the Vespucio dataset from the Chilean toll system, integrated into a cloud data warehouse on Snowflake. It includes comprehensive information on vehicles, passengers, billing, and payment, facilitated by Ardian’s IT and Data Science team. Vespucio has also given additional information on credit scores by providing the Equifax dataset which is also on Snowflake.

The secondary channel involved open-source macroeconomic data, focusing on four areas: overall economic health, inflation, household conditions, and urban transportation. This data was collected from publicly accessible datasets provided by organizations such as The World Bank, International Monetary Fund (IMF), Central Bank of Chile, National Statistics Institute of Chile, OECD Stats, and CEIC data, primarily in CSV or XLSX formats.

**2.2. Quantitative Data**

**2.2.1 Open Source Macroeconomic Data**

As defined in the previous section, we identified 4 categories of macroeconomic data with various geographic scopes and timeframes:

**(1) Overall Economic Health:** For this section, GDP indicators and the unemployment rate were 2 essential factors underconsideration. Details are specified as follows:

| **GDP Indicators** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| GDP, current prices (billions of pesos) | Chile | 1997-2023 | Quarterly | Central Bank of Chile |
| GDP, current prices (billions of U.S. dollars; PPP; billions of international dollars) | Chile | 2018-2028E | Annually | International Monetary Fund |
| GDP per capita, current prices (PPP; international dollars per capita) | Chile | 2018-2028E | Annually | International Monetary Fund |
| Real GDP growth (annual percent change) | Chile | 2018-2028E | Annually | International Monetary Fund |
| GDP by economic activity (sectors), current prices, reference 2018 (billions of pesos) | Santiago | 2013-2022 | Annually | Central Bank of Chile |
| **Unemployment** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| Unemployment rate (percent) | Chile | 1980-2028E | Annually | International Monetary Fund |
| Unemployment rate (percent) | Santiago | 1980-2023 | Monthly | CEIC (paid data source) |

**(2) Inflation:**

| **Inflation Indicator** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| CPI, CPI without volatiles and volatile CPI, base 2023=100, index, spliced ​​information | Chile | 1998-2024 | Monthly | Central Bank of Chile |
| Inflation rate, average consumer prices (annual % change) | Chile | 1980-2028E | Annually | International Monetary Fund |
| Inflation, GDP deflator (annual %) | Chile | 1961-2022 | Annually | Chile |
| Implied PPP conversion rate (national currency per international dollar) | Chile | 1980-2028E | Annually | International Monetary Fund |
| Transportation Cost Index (TCI), national coverage - base year 2018, analytical indices | Chile | 2019-2023 | Monthly | The National Statistics Institute of Chile |
| Supermarket sales index (ISUP) at constant prices, index, 2018=100 | Santiago | 2014-2023 | Annually | Central Bank of Chile |

**(3) Household conditions:**

| **Household** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| Annual national disposable income, net saving, lending and borrowing, in millions of Peso | Chile | 2019-2022 | Annually | OECD Stats |
| Household debt, loans and debt securities (% GDP) | Chile | 2002-2022 | Annually | International Monetary Fund |

**(4) Urban Transportation:**

| **Drivers, Vehicles, and Traffic Info** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| Number of driver's licenses processed (by class, type of processing, age, sex, and region) | Chile | 2011-2022 | Annually | The National Statistics Institute of Chile |
| Passage of vehicles through toll plazas and interurban highway gantries, units | Santiago | 2014-2023 | Annually | Central Bank of Chile |
| Passage of vehicles through urban highway portals, units | Santiago | 2014-2023 | Annually | Central Bank of Chile |
| Fleet / Park of taxi, buses, minibus, school bus, units | Santiago | 2014-2023 | Annually | Central Bank of Chile |

**2.2.2 Vespucio Dataset**

The primary database provided by Vespucio contains detailed transaction-level data for each invoice sent to a customer. This comprehensive dataset spans from 2006 to 2024, comprising approximately 482 million records across 39 columns. Each record encapsulates a wealth of information about the transaction, including the date, characteristics and status of the document, and the transaction amount. Additionally, the dataset includes extensive customer-specific details such as the number of vehicles owned, their Unique National Identification Number (RUT), customer type, region of residence, payment dates, and the time lag between the transaction and payment dates. This extensive dataset is on Snowflake and we have leveraged Snowpark ML to perform computations on the large dataset.

**2.2.3. Equifax Dataset**

Equifax, a global entity recognized for its data, analytics, and technology solutions, operates in Chile as a credit reporting agency. It provides credit information and risk management services to a wide range of clients including financial institutions, businesses, and individual consumers. The Equifax dataset specific to Chilean citizens comprises approximately 3 million records distributed across 7 columns. The credit scores in the Equifax data range from 280 to 850, categorized into various segments that reflect the creditworthiness of individuals. The categories, defined in the following table, help in assessing the financial reliability of individuals, in turn enabling more informed decisions regarding credit and allowing them to mitigate risk effectively.

| 300 - 579 | 580-669 | 670-739 | 740-799 | 800+ |
| --- | --- | --- | --- | --- |
| POOR | FAIR | GOOD | VERY GOOD | EXCELLENT |

## **2.3. Qualitative Data**

Our project exclusively utilizes structured data from the Vespucio dataset, Equifax credit information, and macroeconomic indicators. Given the quantitative nature of our objectives, qualitative data collection was not part of our efforts.

**2.4. Data Processing and Analysis**

**2.4.1 Open Source Macroeconomic Data**

We processed open-source macroeconomic data for correlation analysis by downloading it in CSV and XLSX formats and using AWS for processing. We formatted data on EasyMorph by converting it from horizontal to vertical format, matching dates with corresponding data descriptions and values. The reporting date was standardized across various data types, with an additional date column for consistency, crucial for integrating with the Vespucio dataset. We meticulously preprocessed data to remove empty fields and redundancies, appending an extraction date to each table. Lastly, we used EasyMorph connectors for bulk export to the Snowflake environment, preparing the datasets for further correlation analysis.

**2.4.2 Vespucio and Equifax Data**

For the predictive model using Vespucio and Equifax data, we undertook extensive preprocessing and feature engineering to optimize the dataset for subsequent analysis.

First, the primary dataset was **merged** with Equifax scores based on each customer's RUT, a unique Chilean identifier, ensuring each customer's credit score was accurately aligned with their invoices. This merged dataset was then uploaded to Snowflake to facilitate centralized analysis.

Then, in **refining the dataset**, we filtered invoices from 2021, 2022, and 2023, focusing specifically on transactions labeled as 'Peage' (toll collections) to align with project objectives. During data cleaning, we removed ten columns that served as unique identifiers, such as invoice IDs and billing numbers, to prevent model overfitting. Additionally, any columns with over 70% missing values were excluded, and approximately 4% of outlier records were identified and removed using the Inter-Quartile Range (IQR) method, specifically targeting anomalies in 'DiasCompensacion' and 'ImportePesos'.

**Feature engineering** was a critical component of our preprocessing efforts. We introduced new features such as 'Month\_invoiced' and 'Day\_of\_the\_week\_invoiced,' both derived from the invoice date, along with 'Customer\_lifespan,' calculated from the duration between a customer's first and last appearance in our system.

We also **addressed multicollinearity** by removing eight columns that displayed high correlations with one another. This reduction included columns that redundantly represented codes in both words and numbers, along with three columns tied to regional customer information. Correlations were rigorously assessed using Pearson’s correlation for continuous variables and Chi-squared tests for categorical variables, ensuring that the retained features provided distinct and relevant information for our predictive modeling.

**2.5 Feature Exploration, and Statistical Analysis**

**2.5.1 Macroeconomic Analysis**

To understand the relationships between individual macroeconomic indicators and their influence on default rates, we conducted an extensive feature exploration and statistical analysis. 19 macroeconomic indicators were grouped into four categories: Overall Economic Health, Inflation, Household Conditions, and Urban Transportation. Our aim was to identify the most relevant indicator within each category to accurately explain default rates.

We adopted a series of statistical analyses, primarily using Ordinary Least Squares (OLS) regression. First, we conducted an initial regression analysis by applying univariate OLS to various macroeconomic indicators, treating default rates as the dependent variable. To ensure comparability, each indicator was standardized for scale and normalized. This preliminary assessment helped us understand the initial relationships between indicators and default rates.

To evaluate the reliability of our results, we utilized the Durbin-Watson statistic to detect autocorrelation in the residuals of our regression models, revealing autocorrelation problems in 3 out of the 4 categories. In categories unaffected by autocorrelation (Urban Transportation), we focused on statistically significant indicators and preferred those with monthly or quarterly data to capture detailed trends. For categories impacted by autocorrelation (Overall Economic Health, Inflation, and Household Conditions), we computed the growth rate and applied moving average models to both the growth rate and original variables, with the objective to smooth out short-term fluctuations and highlight long-term trends. To choose the best window size for the moving average, we balanced statistical significance, Durbin-Watson scores, and window size. We carefully managed trade-offs, as a smaller window might show higher statistical significance but less smoothing, while a larger window could over-smooth and lose granularity.

After comparing the original and growth variables, both smoothed, we selected the final indicators that demonstrated the best predictive power while effectively controlling autocorrelation. Our analysis ultimately identified Chilean GDP, Chilean CPI, Household debt, loans, and debt securities (%GDP), and Santiago vehicle passage through toll plazas (units) (yearly data) as the key indicators for our final univariate models.

In the multivariate macroeconomic analysis, the feature exploration method is built upon the insights gained from the univariate analysis. In addition to leveraging the relationships identified between individual indicators and default rates, an Ordinary Least Squares (OLS) regression incorporating all variables was conducted as an initial test to evaluate their collective statistical significance. However, the complete model generated NaN outputs, signaling issues to be further tested. This necessitated a subsequent refinement of the approach, selectively including variables based on their individual contributions and interactions observed. Further visualization techniques will be employed to aid in selecting the most appropriate modeling methodologies moving forward.

**2.5.2 Predictive Model**

Our EDA was focused on exploring variables that may potentially affect StatusCompensacion, This variable contains three categories: Pagada (paid), Impaga (Unpaid), and Castigo (Canceled). After initial review, we decided to carry out EDA for the following variables, which were viewed as most likely to have significant impacts on payment behavior (the graphs for these are in the appendix):

| **Variable** | **Hypothesis** | **Notes** |
| --- | --- | --- |
| Previous StatusCompensacion | If an individual is a consistent payer, they are more likely to pay again, and if they are consistent defaulters, they may default again; customers who have been using free-flow for a longer time may have a higher possibility to be a consistent payer. |  |
| Transaction Amount | The higher the fee, the more likely for people not willing to make payment, as it is too expensive. | Amount can be negative if it is prepayment |
| Document Type | We will expect the collections rate for certain types of orders (such as ordered labeled as “bad”) to be significantly lower | Bad order refers to orders labeled as canceled or evaded |
| Product Type | These tagged drivers should behave better than infractor and evaders in making payment | Two types: Tag, which are cars with sensors, and Infractor |
| Payment Delay | People willing to make payment will be more likely to concentrate in a lower payment delay, and the longer the delay, the higher the possibility for not making payment | Recorded as NaN for those who have not paid |
| Equifax Score | For individuals making payment, they should be more likely to have a higher average score |  |

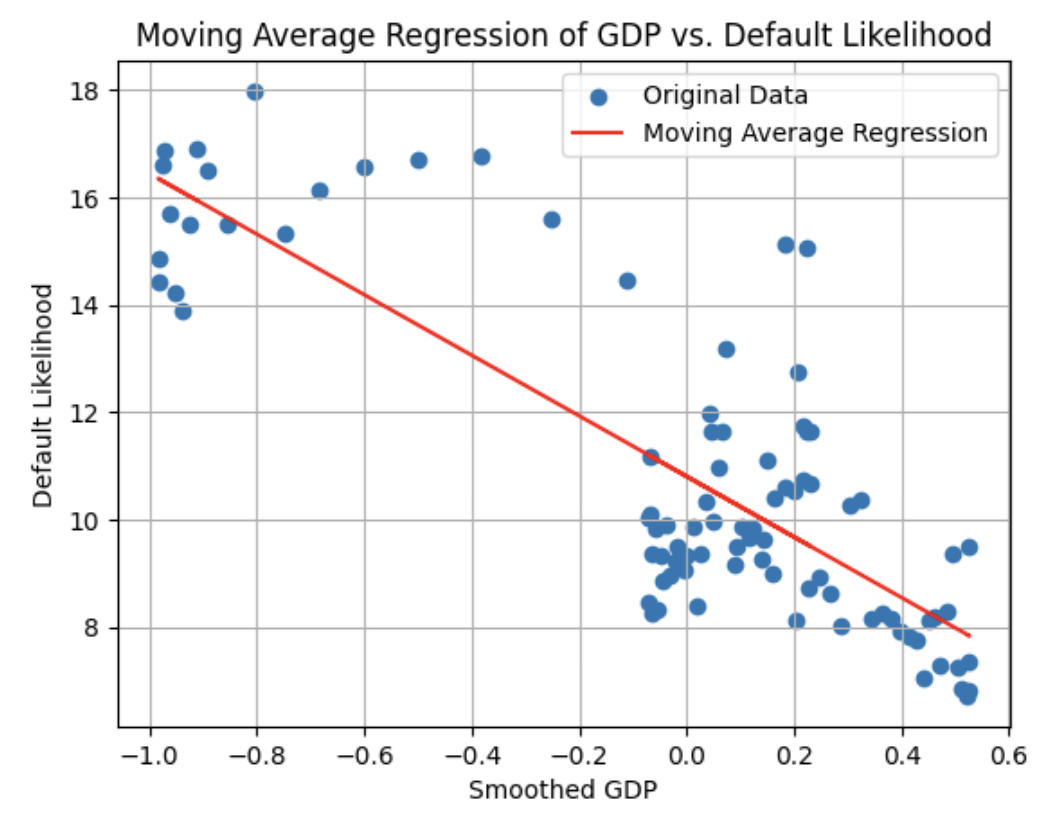
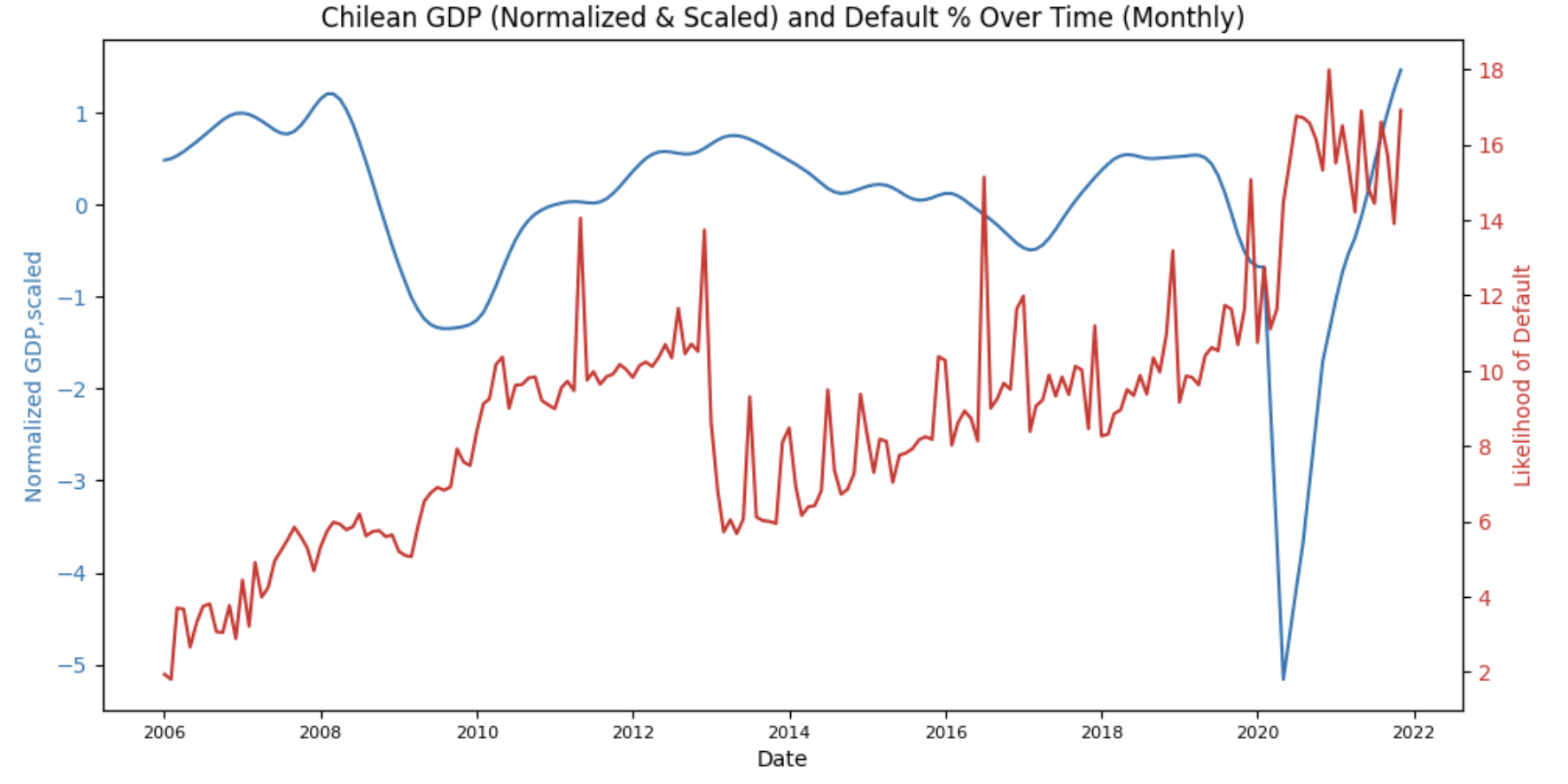
**2.6. Data Visualization**

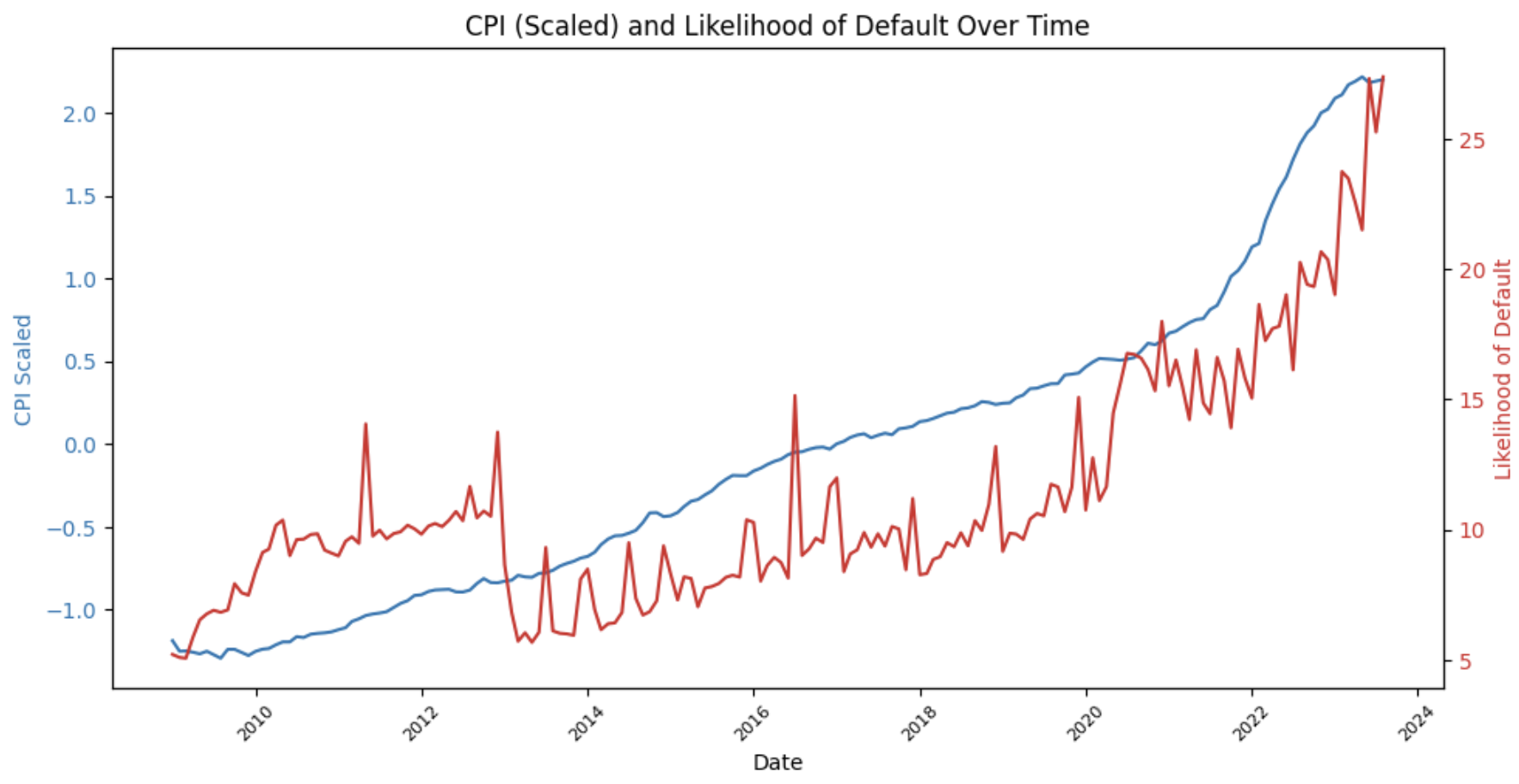
**2.6.1 Macroeconomic Analysis: Univariate**

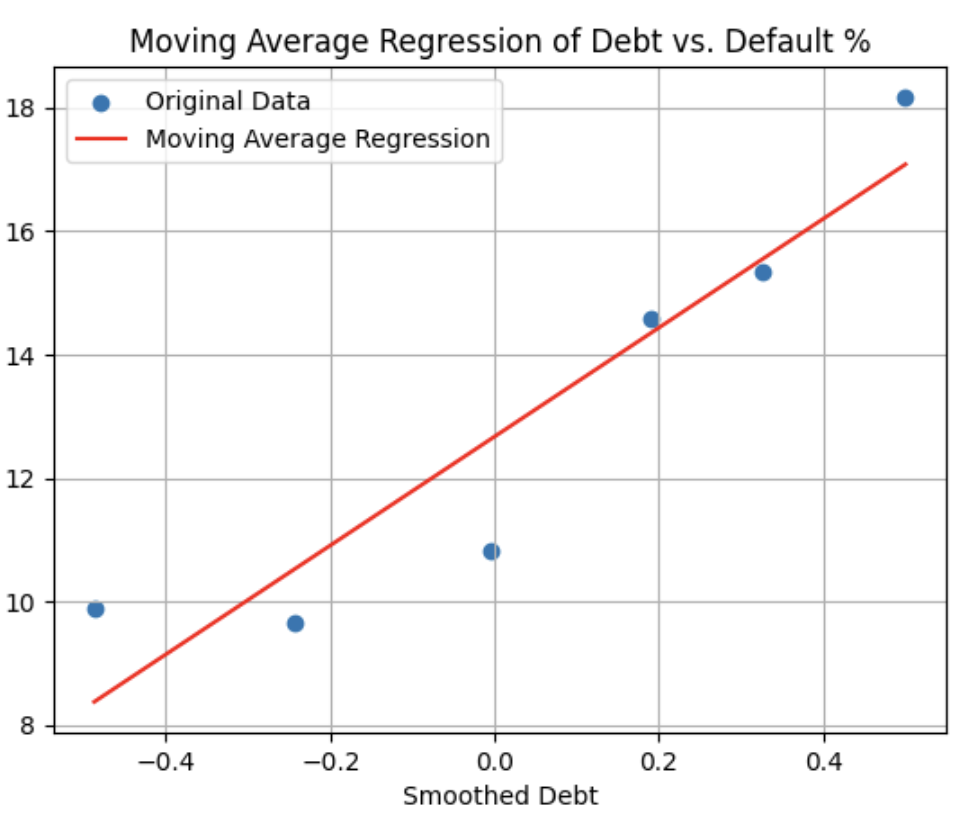
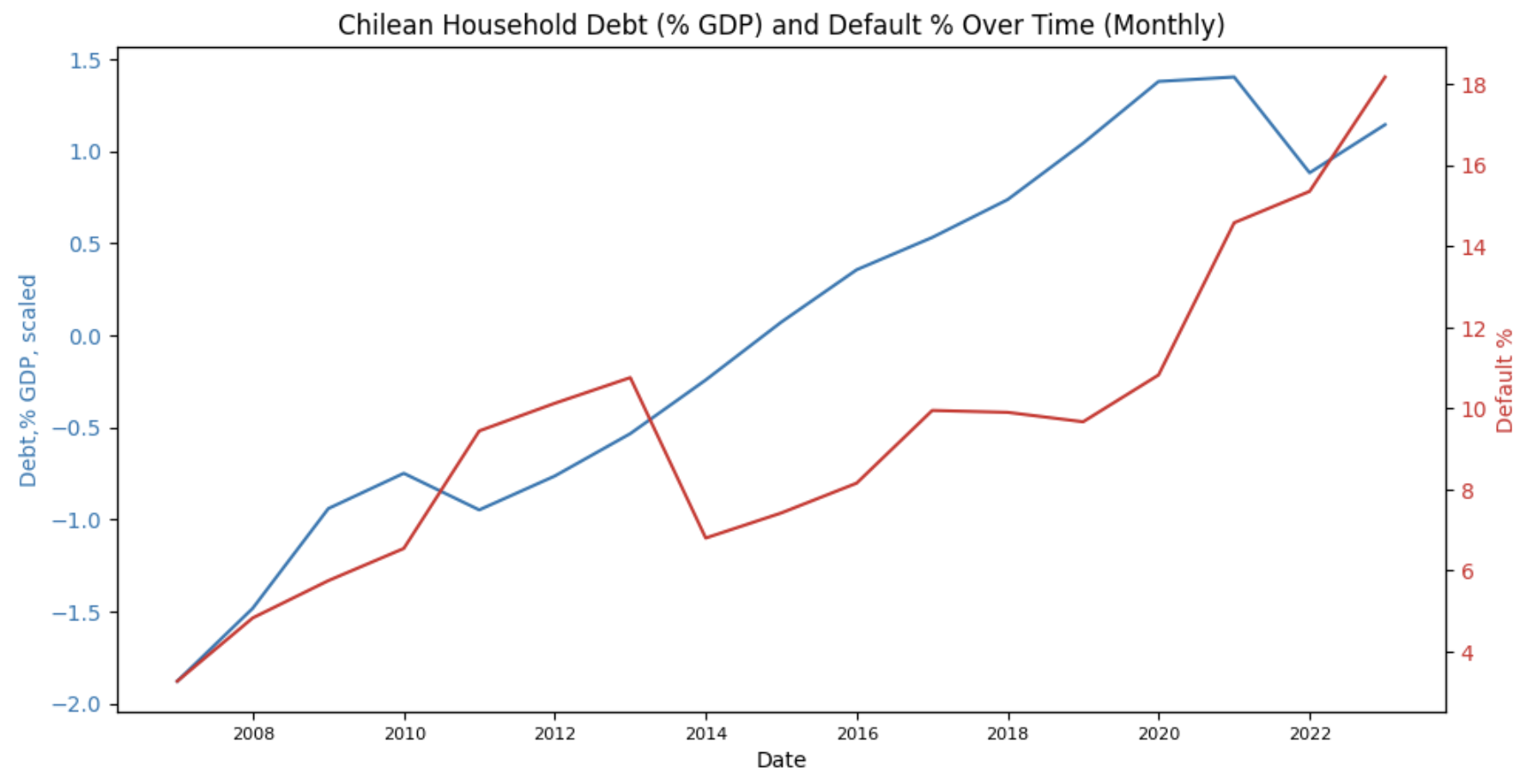
To capture the trends and fluctuations for each selected macroeconomic indicator, we plotted the original variable values against default rates. These visualizations allowed us to observe seasonal and long-term trends in economic data while identifying divergences that revealed anomalies in payment behaviors. For instance, GDP trends were compared to default rates to detect periods where changes in economic growth significantly impacted toll revenue, the direction and strength of impacts, and their lagged effects.. Similarly, inflation, household debt, and urban traffic patterns were plotted to highlight their respective influence on toll collection.

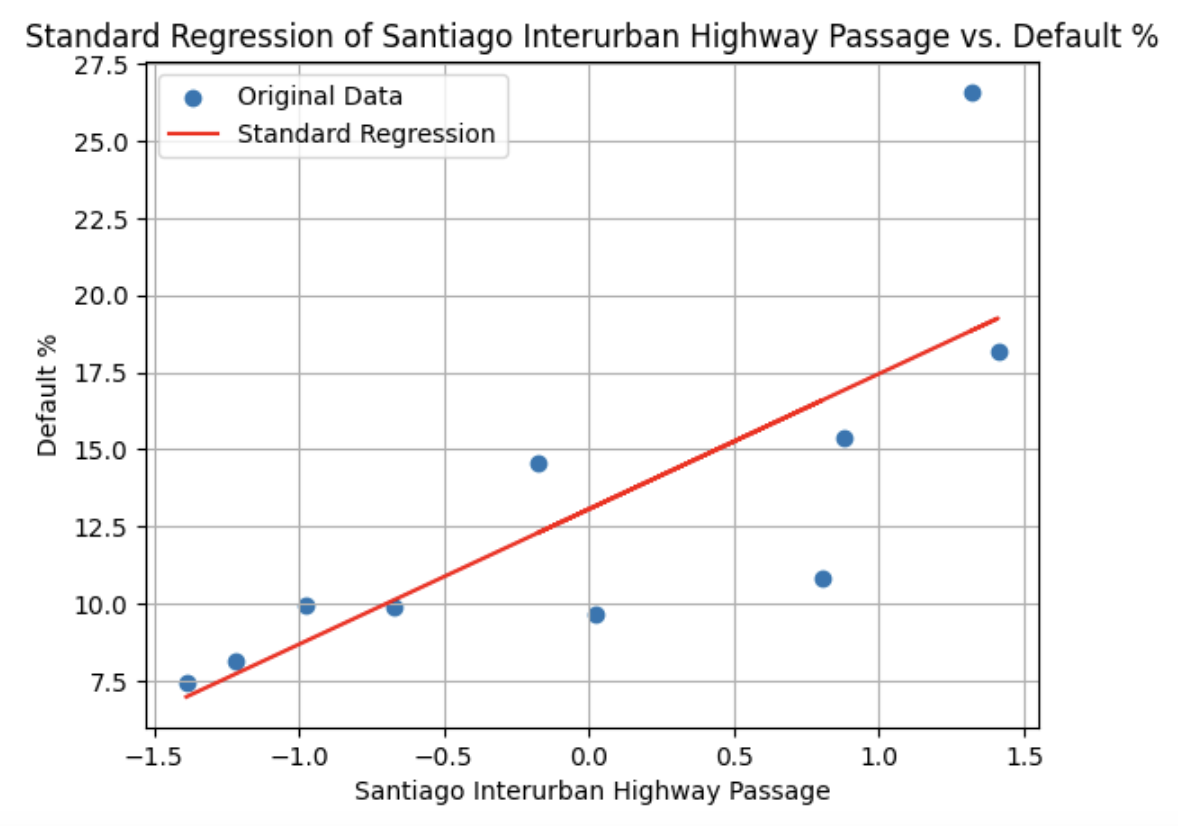
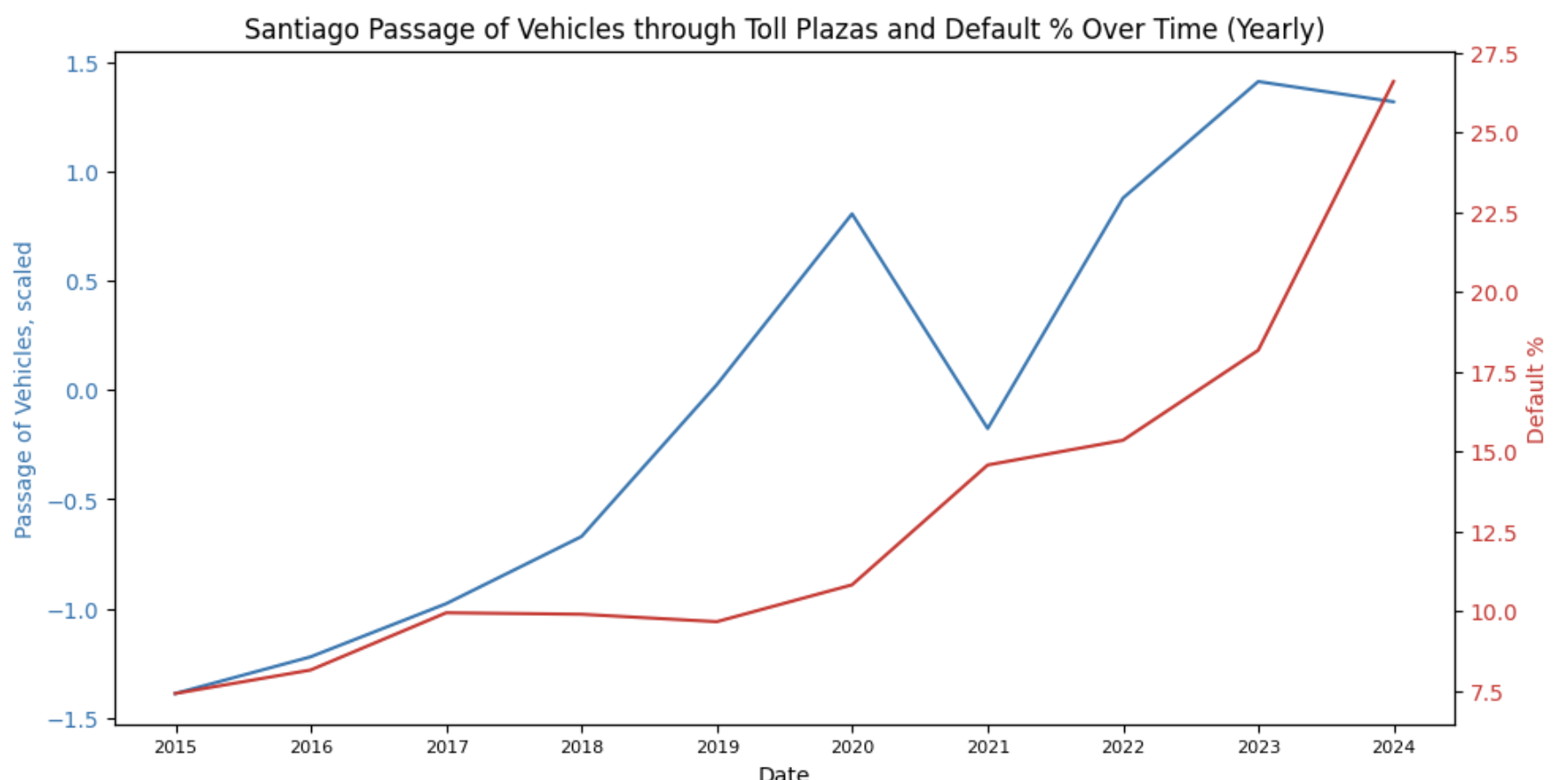
To further analyze the strength of our predictive models, we plotted the original data points alongside the regression line, illustrating how well the model aligns with real-world data. These visualizations exposed any weaknesses in the model's explanatory power, helping us identify outliers and economic factors that warranted additional investigation. For example, periods where the regression line deviated significantly from actual data points highlighted potential external factors affecting payment behaviors.

The visualizations helped us recognize economic indicators' distinct patterns and their relationship to default rates, and offered our audience a clear understanding of the models' predictive strengths and the underlying economic dynamics.







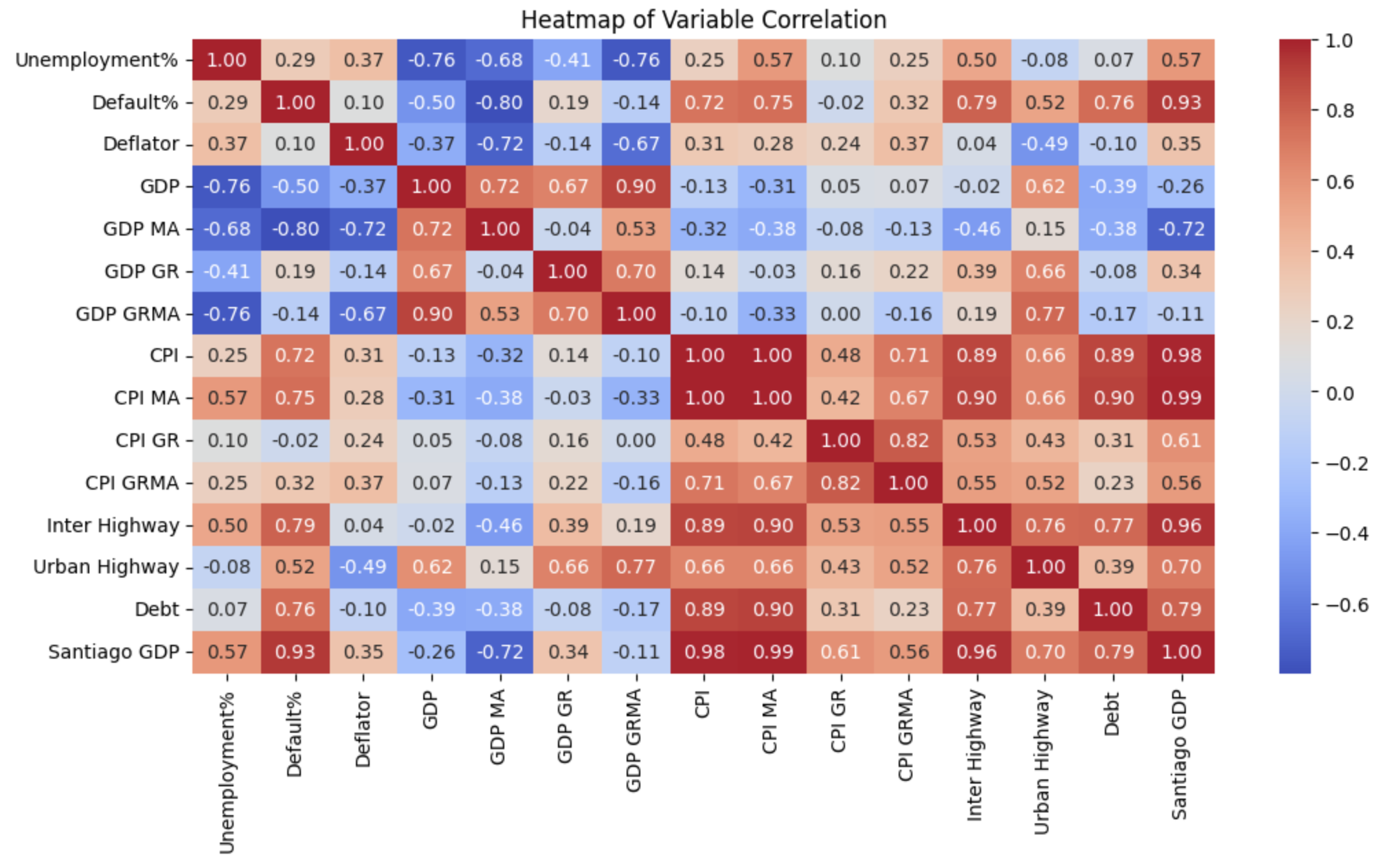
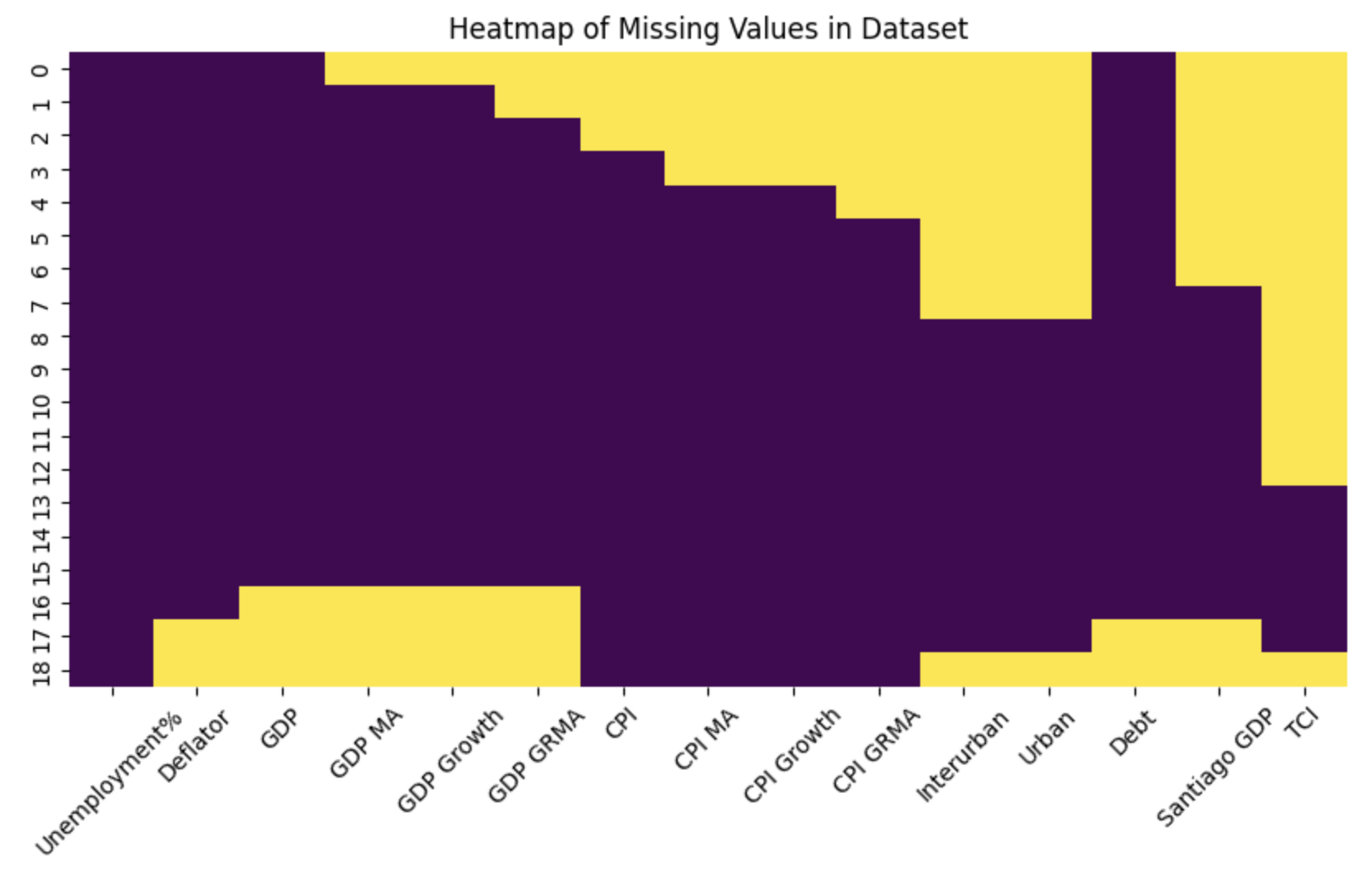


**2.6.2 Macroeconomic Analysis: Multivariate**

To explore the dataset comprehensively and understand the relationships between multiple macroeconomic features, two main visualization techniques were employed: a heatmap of missing values and correlation matrices. Each method serves a distinct purpose and provides insights that are crucial for robust statistical analysis and effective model selection.

The heatmap of missing values is a crucial first step in the data processing phase. It visually represents the structure and completeness of the dataset across different macroeconomic variables. Each row in the heatmap corresponds to a specific observation (yearly), while each column represents a variable. The heatmap outlines the dataset structure with a maximum of 18 observations and 15 variables, with a significant amount of missing values and inconsistencies across variables. Therefore, one necessary step before conducting multivariate analysis is feature selection in order to limit the number of independent variables so there are enough mutual years of observation to be considered as dependent variables. The heatmap also provides some information about which variables to be eliminated based on the number of missing data. For example, TCI (Transportation Cost Index) would be removed from building the model due to its limitation of observations available.

A correlation matrix, visualized through a heatmap, is also a powerful tool to explore data before multivariate correlation analysis, given the nature of macroeconomic indicators that cause natural intercorrelations which can bias regression models. Each cell in the matrix provides a correlation coefficient between two variables, indicating the strength and direction of their linear relationship. The correlation matrix confirms that feature selection is a necessary process before deciding on the multivariate regression model since there are high correlations between many of the variables. The correlation matrix could also be one method in feature selection to determine final indicators.



Overall, the heatmap of missing values and correlation matrix provides insightful information for modeling. They collectively suggest that the first step of building the multivariate regression is feature selection, eliminating the number of independent variables in the model before doing the backward selection. The number of observations and multicollinearity are two major issues to be considered in the later selection and modeling process, in addition to the issues identified in the univariate section.

**2.6.3 Predictive Model**

Our conclusions for the predictive EDA are heavily based on visualizations of the relationships between selected variables and the StatusCompensacion (payment status). In order to keep this section of the report concise, we will present only our conclusions below, with detailed graphs provided in the appendix.

* **Previous StatusCompensacion (Table 1):** Most customers show strong consistency in their payment behavior. By taking the average of the percentile of consistent payment behavior for each RUT individual, we collected an initial consistency of 97.94%, suggesting that for each individual, it is very likely that they will follow their previous payment behavior.
* **Transaction Amount (Graphs 1-3):** We first segmented samples into eight equally sampled bins, the cuts being 2, 8, 81, 409, 662, 1674, and 4469 pesos. At first glance (Graph 2), the transaction amount doesn’t seem to significantly affect the collection rate, and in fact the lowest payment percentile occurred in the bin ranging from 9 to 81 pesos, and the bin ranging from 82 to 409 pesos. Yet, we also noticed that for the range from -9 to 409 and the range from 410 to maximum, there respectively existed our hypothesized trend that higher amounts would lead to lower payment rate.

Taking a closer look at revenues of each bin (Graph 3&4), we realized that there may be regulation methods adopted by Vespucio when the transaction amount is above 410 pesos, as that is when the revenue collected from each bin will start to account for a significant portion of the total revenue. And even with the intervention, from 410 to maximum, we still observed that the collection rate as well as the realized revenue rate was dropping as the transaction amount was increasing. Thus, we considered adding another categorical variable which will be 0 when the Transaction amount is below 410, and 1 when it is above.

* **Document Type (Table 2, Graph 4):** As expected, the collection rate for Boleta Infractora (Evader Receipt) and Castigo Estadistico (Canceled) is significantly lower than other types of documents. But to our surprise, the document type is not very diversified, but highly concentrated in Boleta (Receipt with VAT) and Boleta Exenta (Receipt without VAT).
* **Product Type (Table 3-4):** As expected, the collection rate for Tagged drivers is significantly higher than non-tagged drivers. But, it’s worth noticing that out of 100000 observations, only 238 are identified as non-tagged drivers.
* **Payment Delay (Graph 5):** From the violin plot, it is obvious that for people making payment, they are more concentrated between below 0 days to less than 100 days, while for Castigado, the customers are more evenly distributed between 0 to 1300 days, while the majority is above 500 days. The observation does coincide with our hypothesis.
* **Equifax Score (Table 5, Graph 6):** The boxplot shows that the Pagada group has a significantly higher median than other two groups, with a wider range as well. Taking a closer look into the group mean, we noticed that the Pagada group has a significantly higher average Equifax score than the other two groups. Thus, our hypothesis was rational.

1. **Modeling**

**3.1 Model Selection**

**3.1.1 Macroeconomic Model**

For each macroeconomic indicator category, our objective was to identify a model that could effectively capture the relationship between key economic variables and default rates. We initially explored models using lagged variables for macroeconomic indicators that exhibited significant autocorrelation, which allows us to incorporate the impact of previous time periods to reveal any delayed effects on default rates. Despite testing lags ranging up to 200 periods, we found that autocorrelation remained persistent, leading us to conclude that a more comprehensive strategy was needed. Other methodologies considered included ARIMA models. While these models are well-suited for time series analysis, they introduced unnecessary complexity for our already extensive dataset and did not perform better than simpler models.

For indicators that exhibited autocorrelation, we selected moving average Ordinary Least Squares (OLS) models to smooth short-term volatility and emphasize long-term trends, allowing us to capture predictive signals while mitigating autocorrelation's impact. For indicators that did not show autocorrelation, standard OLS models were chosen to maintain data granularity and capture immediate fluctuations and trends.

The modeling methodology for the multivariate macroeconomic analysis mirrors the approach used in the univariate analysis to maintain consistency and leverage established insights. When indicators demonstrate mutual granularity and permit detailed trend analysis, moving average Ordinary Least Squares (OLS) models are employed to smooth out short-term volatility and focus on long-term trends, thereby enhancing the model's ability to demonstrate default rates effectively while addressing autocorrelation issues. Conversely, for scenarios where data are aggregated on a yearly basis, which restricts the practical window size for moving averages, standard OLS models are utilized. This approach allows for the retention of data granularity, capturing immediate fluctuations and broader trends without the application of moving averages, thus adapting to the limitations imposed by the data's temporal resolution.

**3.1.2 Predictive Model**

The objective of the predictive model was to forecast whether an invoice would be paid, aiding Vespucio in preemptively identifying potential defaults during the invoicing process. The primary goal was not only to predict defaults but also to discern the contributing factors. Initial data exploration, including bivariate and correlation plots and techniques like Select K-Best features enabled us to eliminate several irrelevant variables. We narrowed down our analysis to 14 explanatory variables and one binary target variable (Unpaid: 1 for unpaid, 0 for paid).

We tested five different classification models: a baseline Logistic Regression, XGBoost, LightGBM, Random Forest, and CatBoost, the latter known for its efficacy with categorical data. Tree-based models exhibited similar performance in terms of accuracy, with minor differences in precision and recall metrics. After conducting hyperparameter tuning on the similarly performing models, we selected XGBoost as our final model due to its superior recall rate.

**3.2 Modeling Methodology**

**3.2.1 Macroeconomic Analysis: Univariate**

For the selected macroeconomic indicators showing significant autocorrelation (Chilean GDP, Chilean CPI, and Household debt, loans & debt securities, we employed a moving average Ordinary Least Squares (OLS) model. The moving average was applied to both the growth rate and original variables to reduce the impact of short-term fluctuations and highlight long-term trends.The window size for the moving average was carefully chosen to balance statistical significance, Durbin-Watson scores, and smoothing levels. This optimized window size minimized autocorrelation while preserving the model's predictive power. The moving average OLS model analyzed these smoothed variables to determine their relationship to toll payment default rates.For the indicator that did not show autocorrelation (Santiago vehicle passage through toll plazas), we used a standard OLS model without incorporating a moving average.

We validated the resulting models through comprehensive statistical measures. Key metrics included p-values, F-statistics, and the Durbin-Watson scores to check for model significance and residual autocorrelation. Our models achieved significant p-values, indicating reliable predictive power, and favorable F-statistics demonstrating model fit. We also employed the Jarque-Bera (JB) test to assess residual normality. The models displayed reasonable JB scores, affirming normally distributed residuals.

Skewness and kurtosis statistics were also examined to ensure the models were neither overly skewed nor excessively peaked. This analysis confirmed that residual distributions were relatively symmetrical and lacked extreme outliers, ensuring the final models could accurately explain toll payment default rates and provide valuable insights on toll defaults.

**3.2.2 Macroeconomic Analysis: Multivariate**

In our multivariate macroeconomic analysis, we structured our approach into two key phases: feature selection and backward selection. This structure helped us navigate the complexities of the data and refine our models for accurate predictions of default rates.

During the feature selection phase, we extracted a subset of relevant variables from our broad dataset, guided by insights from initial feature explorations. We employed three different methodologies to optimize the selection process specifically for regression models analyzing correlations with default rates. Backward selection involved starting with all candidate variables and systematically removing the least significant ones based on their p-values. This iterative process was continued until only statistically significant variables remained, thus simplifying the model and improving its interpretability.

* **Method 1. Feature selection by number of observations:** We utilized data segmented by month and year intervals to align the granularity of data with the insights they provided. The objective was to maintain the maximum possible number of relevant variables and observations. Within this framework, two distinct models were tested: one utilizing data with monthly granularity and another segmented on a yearly basis. The monthly model utilized moving averages of the Chilean GDP and CPI, providing a more robust dataset, whereas the yearly model incorporated ten variables and faced challenges with multicollinearity, which hindered its effectiveness even after applying the backward selection process.
* **Method 2. Feature selection by correlation:** We focused on refining our variable selection for the regression model by employing a correlation-based feature selection strategy. We established a correlation threshold of 0.5, aiming to identify variables that not only have a strong relationship with default rates but also exhibit low interdependence with each other. Initially, this approach led to the selection of Santiago GDP, Chilean GDP, and Chilean GDP Deflator as key variables. To further optimize our model, we applied backward selection, which helped us refine the selection down to the most significant predictors: Santiago GDP and Chilean GDP Deflator. This method ensured that our final model was both streamlined and potent in predicting default rates.
* **Method 3. Feature selection by univariate performance:** The third method we utilized focuses on univariate performance, where indicators are selected based on their effectiveness in univariate regression models. Initially, we developed two distinct models: the first incorporated moving averages of CPI and GDP, both aggregated on a yearly basis for multivariate regression, to account for fluctuations and enhance the stability of the analysis. This model began with variables like Chilean GDP Moving Average, Chilean CPI Moving Average, Household Debt, Loans, and Debt Securities (%GDP), and Santiago Passage of Vehicles through Toll Plazas (Unit). Using backward selection, we refined this model to focus on the Chilean GDP Moving Average and Santiago Passage through Toll Plazas.
* **Method 4. Feature selection by univariate performance without moving averages:** To address the potential limitations posed by moving averages aggregated in years, we also crafted a second model without moving averages. Starting with a similar set of variables—Chilean GDP, Chilean CPI, Household Debt, Loans, and Debt Securities (%GDP), and Santiago Passage through Toll Plazas—we applied backward selection once more. This process resulted in a streamlined model that featured Chilean GDP, Chilean CPI, and Household Debt, Loans, and Debt Securities, which proved to be significant predictors without the smoothing effect of moving averages. This methodological variation allows us to assess the impact of data smoothing techniques on the accuracy of our economic predictions.

**3.2.3 Predictive Model**

In our predictive model, we used a dataset comprising 14 explanatory variables and one target variable to analyze payment defaults. From an initial dataset size of over 102 million records, we sampled around 1.02 million rows to balance manageability with computational efficiency.

Explanatory variables, selected through the aforementioned approach, included ImportePesos (Transaction Amount), CantidadCuentasContrato (Number of Contracts), CantidadVehiculos (Number of Vehicles), Score Equifax (Credit Score), Customer\_lifespan, Inhabilitado (Tag User Disabled), Concesionaria (Highway Concession Type), TipoCliente (Type of Client), TipoDocumento (Type of Document), TipoProducto (Tag/Infractor), Day\_of\_week\_invoiced, MaxCategoriaVehiculo (Vehicle Category), and CIUDAD (City). The target variable, 'Unpaid', distinguishes between paid and unpaid invoices, the latter also including canceled transactions due to long periods of default.

Given the significant class imbalance in the target variable, with a 13% unpaid to 87% paid ratio, special attention was needed during model training to ensure accuracy across both classes. Our approach included several preprocessing steps to optimize the data, thus addressing the class imbalance and also enhancing predictive accuracy:

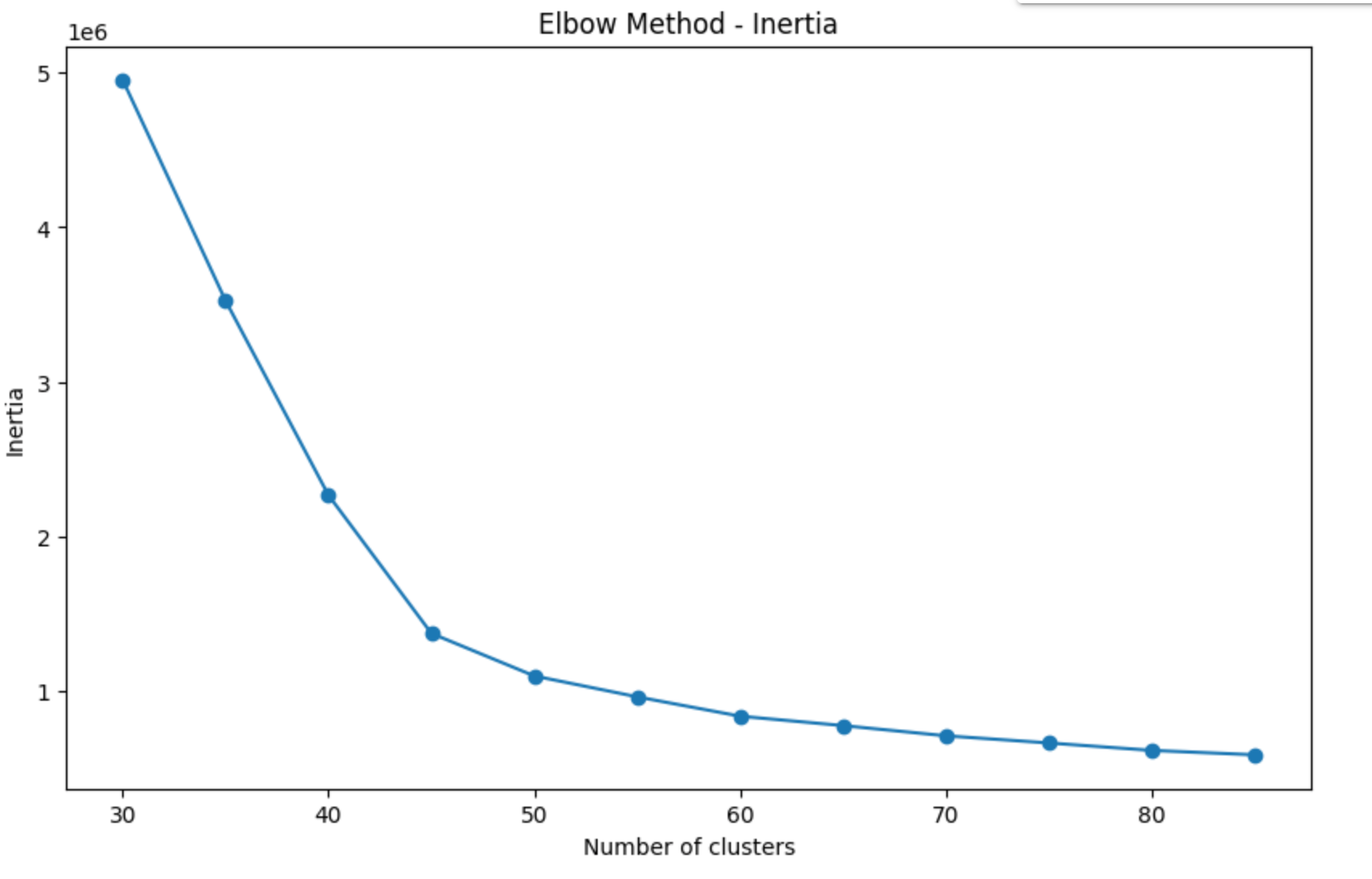
* **Feature Scaling:** Numerical features were standardized and normalized to have mean zero and unit variance, which aids in faster convergence during model training.
* **Encoding Categorical Variables:** One-hot encoding was applied to categorical variables to convert them into a machine-learning-friendly format, ensuring that the importance of these features was accurately captured without implying any ordinal relationship.
* **Train-Test Split:** We divided the dataset into an 80:20 ratio for training and testing, ensuring robust training and a reliable evaluation of model performance.
* **Model Training and Evaluation:**We trained multiple models including Logistic Regression, XGBoost, Random Forest, LightGBM, and CatBoost on the 80% training data. Each model’s performance was evaluated by its ability to predict defaults accurately, using a decision threshold of 0.5 to classify transactions into 'default' or 'no default.'
* **Hyperparameter Tuning:** We focused on fine-tuning the XGBoost and LightGBM models by adjusting parameters such as the number of estimators, tree depth, and learning rate. Although hyperparameter tuning improved the models' performance slightly (about a 0.7% increase), the enhancements were marginal. An exhaustive search with GridSearchCV identified the best parameter settings, suggesting that creating more pertinent features might yield more substantial improvements than hyperparameter adjustments by better capturing underlying data complexities.

**3.2.4 Clustering**

To enhance our predictive model and address the heterogeneities in our dataset and customer base, we implemented a clustering approach. The aim was to segment our customers into distinct groups, characterized by similar traits and behaviors, to enable the development of more precise, group-specific predictive models.

Our first step was to select the most relevant features for clustering. We focused on variables closely related to default, categorized into geographical, individual-level, and continuous variables. Graphs for variable relationships to default are provided in the Appendix, Graphs 7-15.

* **Geographical Variables:** We observed significant differences in default rates across regions, cities, and counties. By incorporating these variables, our clustering algorithm could identify regions with similar risk profiles, facilitating the creation of localized predictive models.
* **Individual-Level Variables:** These included the type of customer (natural or legal entity), the number of contracts, and specific attributes related to highway concessions. These variables helped us distinguish between customer groups with different default rates.
* **Continuous Variables:** We analyzed continuous variables for their correlation with default rates through logistic regression. Important variables included transaction amount (ImportePesos), number of vehicles (CantidadVehiculos), and Equifax score. Notably, higher transaction amounts correlated with increased default risks, while a higher number of vehicles generally indicated lower default probability, especially for entities owning over 100 vehicles—likely due to their legal constraints and lower risk profile.

We initially applied the elbow method to determine the optimal number of clusters by evaluating the change in inertia. While the method suggested 40 clusters might minimize intra-cluster variance, aiming for practical application, we opted to form 10 distinct clusters. This number allowed us to achieve sufficiently differentiated groups, each reflecting notable differences in default rates, thus making our clustering approach manageable yet effective.

1. **Results and Insights**

**4.1 Macroeconomic Model: Univariate**

Full Model results are provided in the Appendix, tables 6-9, with brief descriptions provided below. The results suggest that macroeconomic factors like GDP, CPI, and household debt, and urban transportation trends, have strong and often lagged effects on toll payment default rates, highlighting economic relationships that can help stakeholders anticipate and mitigate toll revenue losses effectively:

* The OLS model using a **moving average of GDP** **(window = 33)** achieved an R-squared of 0.582, indicating that it explains 58.2% of the variation in toll payment defaults. The model's F-statistic of 77.83, with a p-value of 3.50e-12, confirms the model's overall significance. While the Durbin-Watson score of 0.682 suggests some autocorrelation in the residuals, the Jarque-Bera test shows normally distributed errors (p = 0.197), highlighting a well-fitting model. The coefficient for GDP\_MA is statistically significant (p < 0.001), emphasizing a strong negative relationship between GDP and default likelihood.
* For the **CPI moving average model (window = 31)**, the R-squared value was 0.848, explaining 84.8% of the variation in default rates. The high F-statistic of 257.0 (p-value 1.88e-20) underscores the robustness of this model. Residuals displayed moderate autocorrelation (Durbin-Watson score = 1.576), but residuals remained normally distributed (Jarque-Bera p = 0.533). The CPI coefficient is significant (p < 0.001), demonstrating a direct positive relationship between inflation and default rates.
* The **moving average model for household debt (window = 12)** resulted in an R-squared of 0.874, explaining 87.4% of the variation in default likelihood. With an F-statistic of 27.76 (p-value 0.00622), this model is statistically significant. The Durbin-Watson score (1.652) and Jarque-Bera p-value (0.837) indicate manageable residual autocorrelation and normally distributed errors, respectively. The household debt coefficient is significant (p < 0.01), highlighting a positive relationship between household debt and default likelihood.
* The **standard OLS model with the Santiago Passage Index** yielded an R-squared of 0.624, explaining 62.4% of the variation in toll payment defaults. With an F-statistic of 13.25 (p-value 0.00658), this model is significant. The Durbin-Watson score of 1.517 and the Jarque-Bera test (p = 0.681) suggest residuals are normally distributed and have low autocorrelation. The index coefficient is significant (p < 0.01), showing a positive correlation between toll plaza traffic and default rates.

**4.2 Macroeconomic Model: Multivariate**

| **Model** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| --- | --- | --- | --- | --- |
| **Independent Variables** | **1. Chilean GDP, Moving Average**  **2. Chilean General CPI, Moving Average** | **1. Santiago GDP**  **2. Chilean GDP Deflator** | **1. Chilean GDP, Moving Average**  **2. Santiago Interurban Toll Passage** | **1. Chilean GDP**  **2. Chilean General CPI**  **3. Chilean Household Debt** |
| **Granularity** | Monthly | **Yearly** | **Yearly** | Yearly |
| **Coefficients** | Const: 10.7690 (P>|t|=0.000)  GDP MA: -1.7772 (P>|t|=0.000)  CPI MA: 1.8592 (P>|t|=0.000) | **Const: 11.0637 (P>|t|=0.000)**  **Santiago GDP: 2.8884 (P>|t|=0.004)**  **Deflator: 1.6029 (P>|t|=0.004)** | **Const: 10.6389 (P>|t|=0.000)**  **GDP MA: -2.0007 (P>|t|=0.000)**  **Interurban Toll: 1.2035 (P>|t|=0.002)** | Const: 14.6659 (P>|t|=0.000)  GDP: -0.8625 (P>|t|=0.049)  CPI: 9.1599 (P>|t|=0.001)  Debt: 4.9070 (P>|t|=0.010) |
| **Model Fit** | Adj. R-squared = 0.702  F-statistic = 153.0 | **Adj. R-squared = 0.952**  **F-statistic = 89.66** | **Adj. R-squared = 0.977**  **F-statistic = 148.7** | Adj. R-squared = 0.740  F-statistic = 12.38 |
| **Diagnostic Tests** | Durbin-Watson = 1.202  Jarque-Bera = 31.775  Cond. No = 1.84 | **Durbin-Watson = 1.726**  **Jarque-Bera = 1.478**  **Cond. No = 1.61** | **Durbin-Watson = 2.724**  **Jarque-Bera = 0.784**  **Cond. No = 1.96** | Durbin-Watson = 1.278  Jarque-Bera = 1.068  Cond. No = 8.75 |

Based on our modeling methodologies, we developed four distinct models, each utilizing a different set of variables to interpret various aspects of the macroeconomy. The regression results are provided in the appendix, tables 10-13, and an overview is provided below:

**Model 1** assessed the monthly data on Chilean GDP and CPI moving averages, revealing a negative correlation with GDP and a positive correlation with CPI. This model achieved an adjusted R-squared of 0.702 and an F-statistic of 153.0, indicating strong explanatory power. However, a Durbin-Watson statistic of 1.202 and a high Jarque-Bera value pointed to possible autocorrelation and issues with the normality of residuals.

**Model 2** focused on yearly data, examining Santiago GDP alongside the Chilean GDP Deflator. Both variables showed a positive correlation with the default, and the model's robustness was highlighted by an adjusted R-squared of 0.952 and an F-statistic of 89.66. The Durbin-Watson statistic of 1.726 suggested slight autocorrelation, but normally distributed residuals.

**Model 3** also utilized yearly data, incorporating Chilean GDP moving averages and Santiago Interurban Toll Passage data, showing a negative correlation with GDP and a positive correlation with toll data. It achieved an exceptional adjusted R-squared of 0.977 and an F-statistic of 148.7, with very low autocorrelation indicated by a Durbin-Watson statistic of 2.724.

**Model 4** analyzed the impact of Chilean GDP, CPI, and household debt without using moving averages, identifying a negative correlation with GDP and positive correlations with CPI and household debt. The model achieved an adjusted R-squared of 0.740 and an F-statistic of 12.38 but showed some autocorrelation with a Durbin-Watson statistic of 1.278.

Models 2 and 3 are particularly notable for their high explanatory power, effectively capturing regional and national economic trends. Model 2 combines Santiago's economic output with the national GDP deflator, reflecting both local and broader economic influences. Model 3 expands this by including data on regional transportation, which adds depth to our understanding of how local mobility metrics interact with economic indicators.

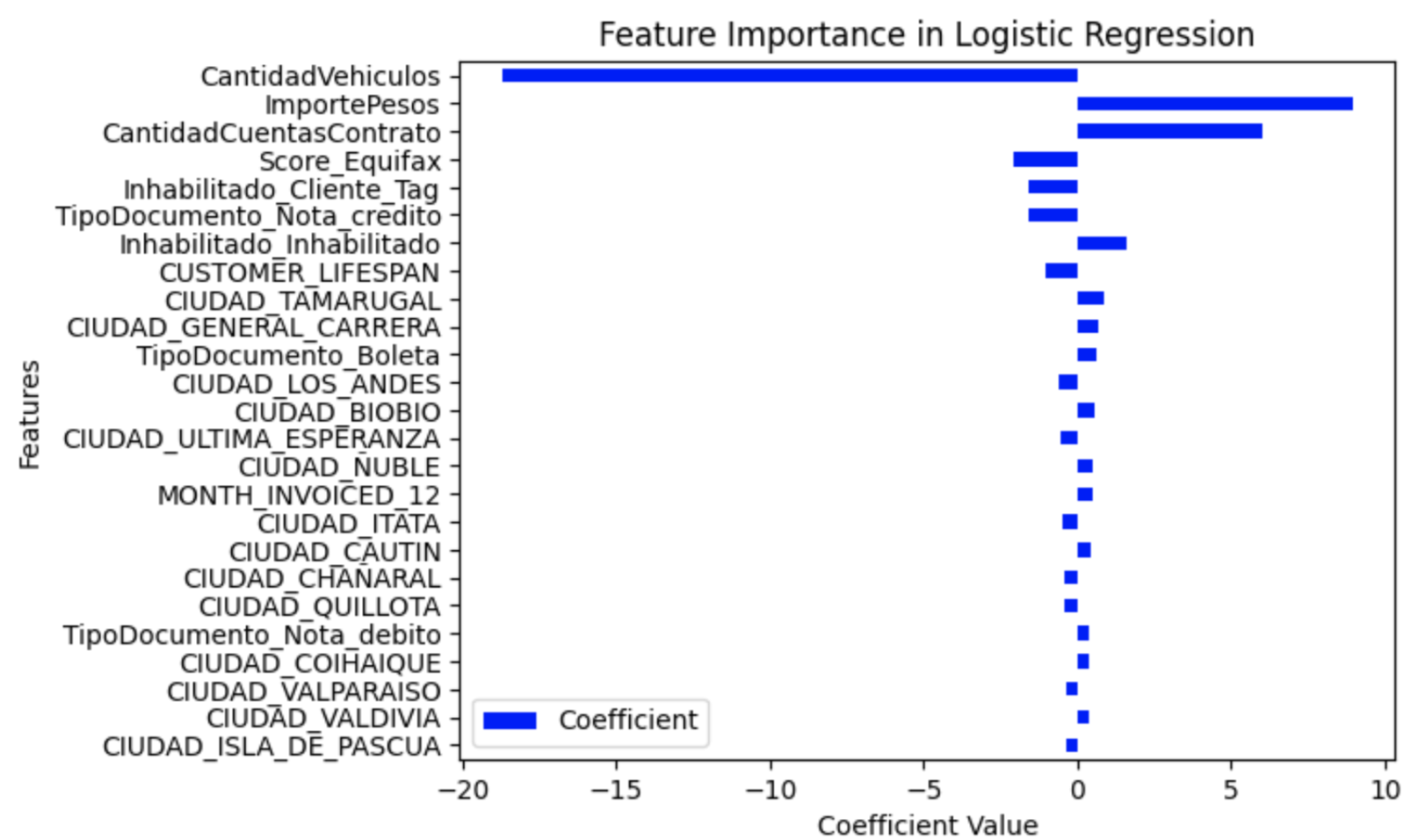
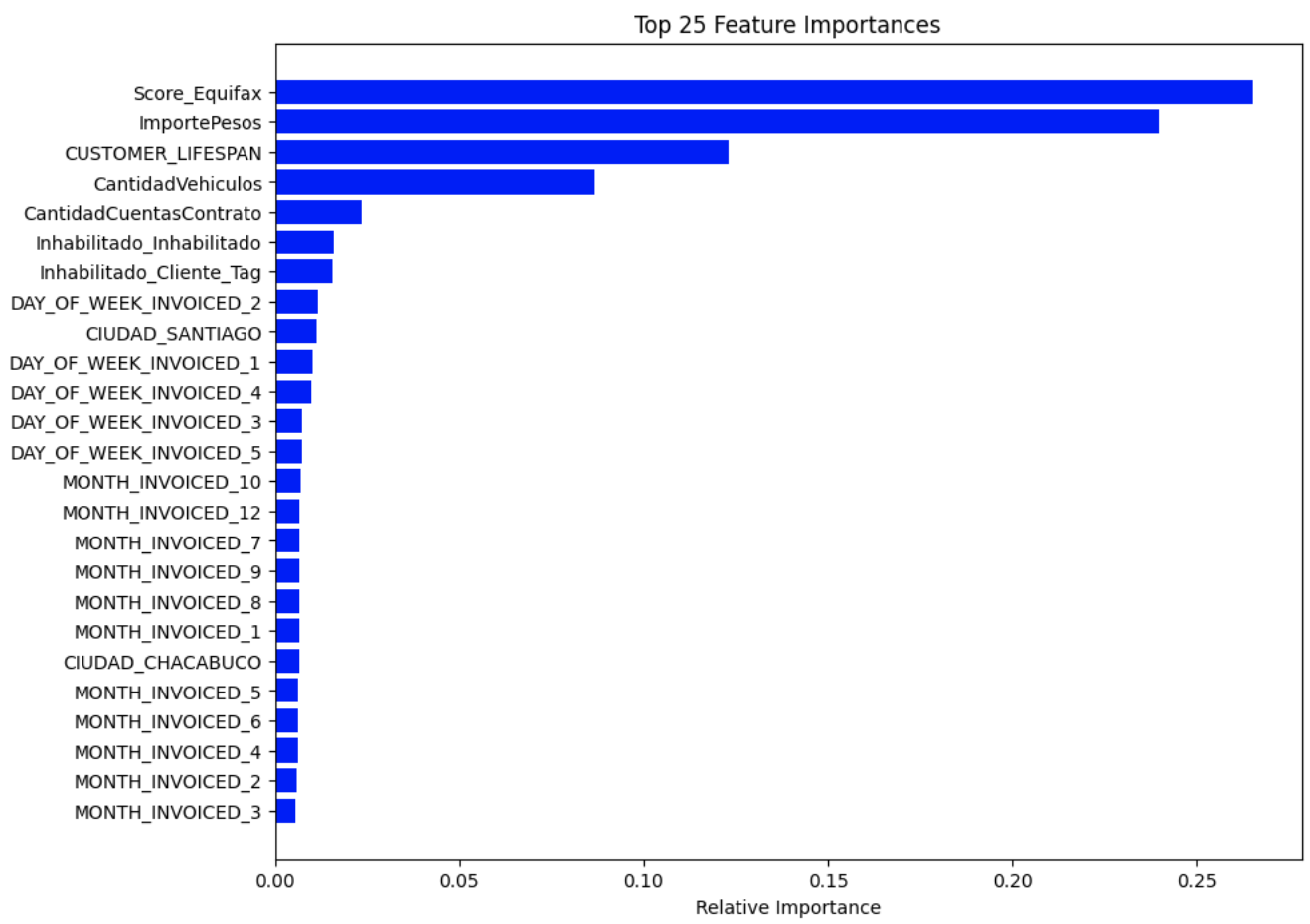
The consistently high intercept across all models imply that default rates tend to remain elevated regardless of economic conditions, indicating systemic issues influencing default rates beyond economic fluctuations. The analysis underscores a general negative correlation between economic growth (as measured by Chilean GDP) and default rates, suggesting that economic expansions tend to reduce default rates. Conversely, inflation metrics like CPI and the GDP deflator are positively correlated with default rates, highlighting the pressures inflation can impose on financial commitments. Integrating both regional and national economic data has enriched our models, allowing for a nuanced interpretation of economic disparities and their impact on financial stability. This comprehensive approach enhances our predictive capabilities and the strategic application of economic insights in managing financial risks.

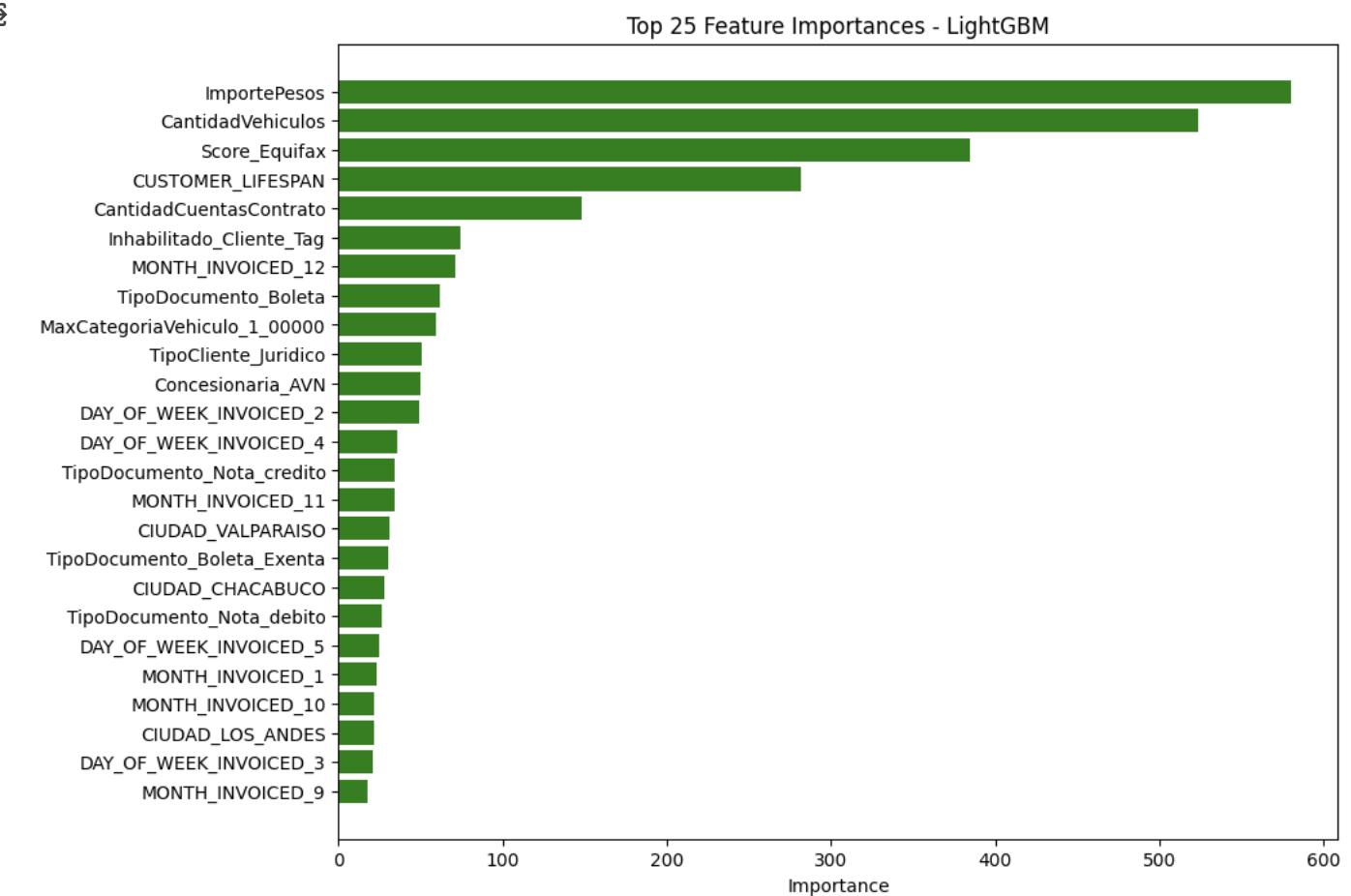
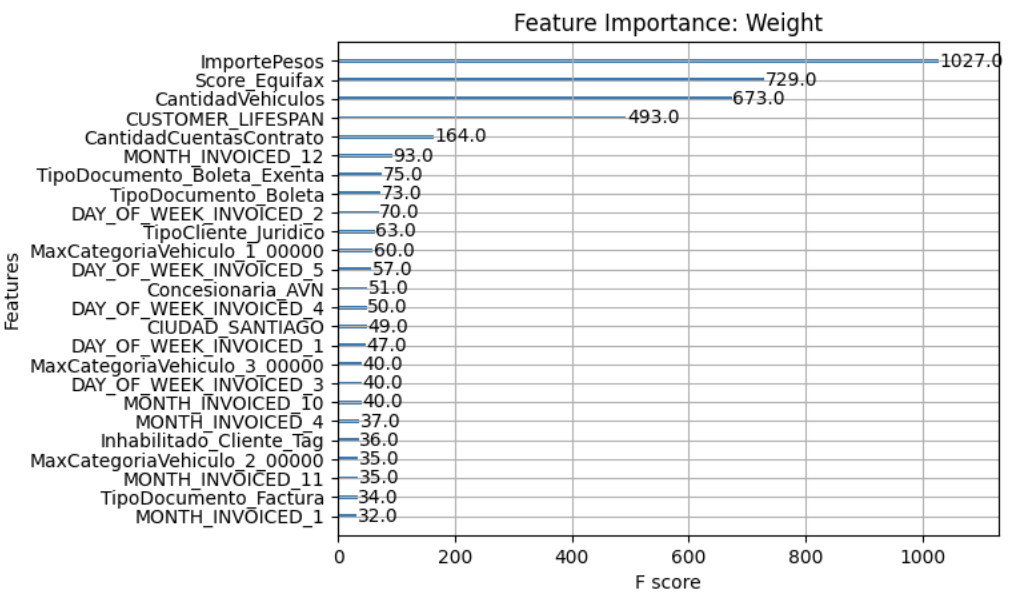
**4.3 Predictive Model**

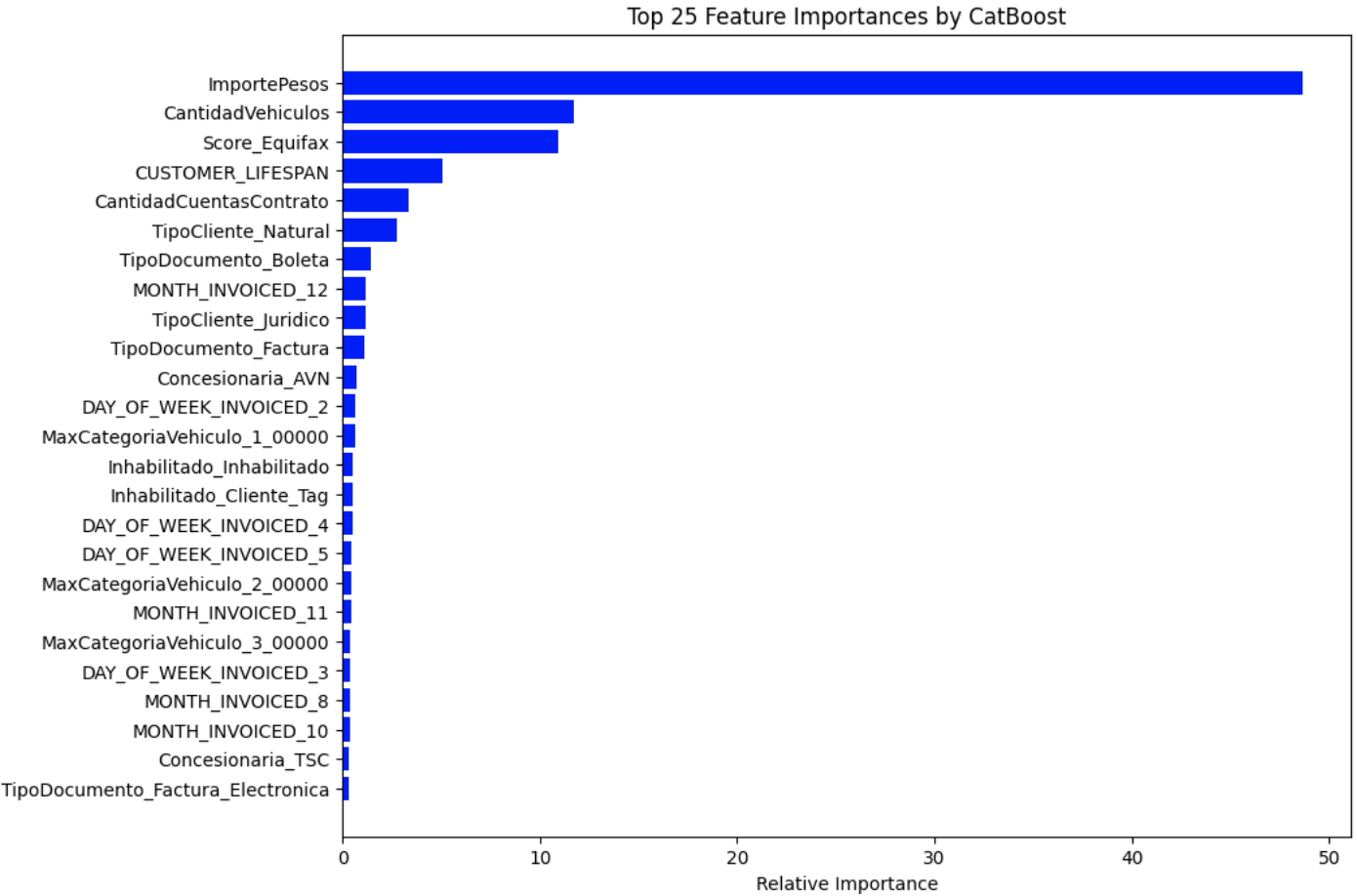
The following table summarizes the accuracy, precision, recall, and F1 score for each of the models used in the project, providing a comprehensive overview of their performance in predicting invoice payment defaults.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- |
| *Logistic Regression* | 88.5% | 72.1% | 5% | 10% |
| *Random Forest* | 88.5% | 57% | 15.5% | 24.4% |
| *XGBoost* | 89% | 70% | 16% | 24.2% |
| *LightGBM* | 88.9% | 72.9% | 11.59% | 20% |
| *CatBoost* | 88.7% | 65% | 15% | 24.9% |

Different models demonstrated varying feature importances in their predictions of invoice defaults. While the overall trend in feature significance remained consistent across models, there were subtle variations in their order of importance. The feature importance plots for each model, detailed below, visually represent these differences and highlight how specific features weigh differently depending on the model used.





**4.3.1 Key observations and findings**:

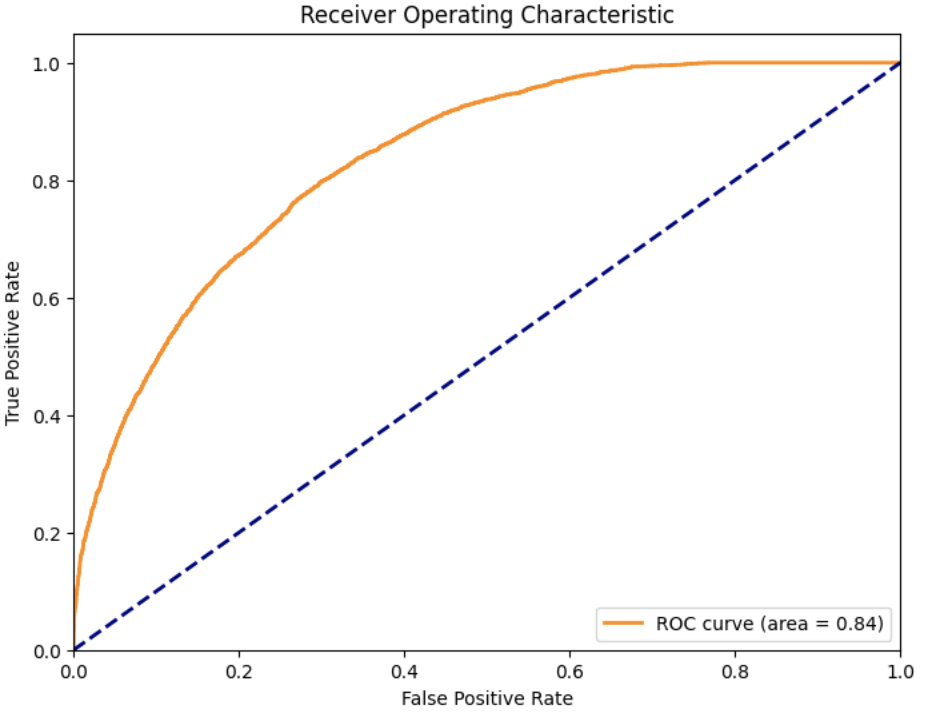
* **Consistency Across Models:** All five models trained exhibited similar patterns in the features deemed important for predicting defaults, though there were slight variations in the rank of importance. Features like ImportePesos (Transaction Amount), Score Equifax (Credit Score), CantidadVehiculos (Number of Vehicles), and Customer Lifespan were consistently highlighted across models, aligning with our initial expectations.
* **Credibility and Transaction amount**: The Logistic Regression model's feature importance plot not only identifies key features but also indicates the direction of their relationship with the likelihood of default. For instance, ImportePesos is positively correlated with default risk, suggesting that higher transaction amounts increase the likelihood of defaulting. Conversely, Score Equifax shows a negative correlation, indicating that a higher credit score, which marks a more credible individual, decreases the likelihood of default.
* **Impact of Vehicle Ownership:** An interesting observation from the models is that an increased number of vehicles (reflected in CantidadVehiculos) typically corresponds to a lower chance of default. This could be attributed to the hypothesis that individuals who own multiple vehicles are generally more financially stable and thus less likely to default.
* **Geographical Variations:** The models revealed that geographical factors also play a significant role in predicting defaults. Residents from certain cities are more prone to defaulting, while others have demonstrated a consistent record of timely payments. This variation suggests that local economic conditions and demographic factors might influence default rates.

**4.3.2 Business Validation**

The importance and direction of these features are not only statistically significant but also align with business logic and validations. The intuitive understanding of these relationships further confirms the reliability of our predictive modeling approach.

While our models are proficient at predicting Paid transactions, the recall for Unpaid invoices—critical for preemptively identifying potential defaults—has room for improvement. The recall rates are not as high as desired, possibly due to the presence of noise and missing information about customers. Additionally, the data imbalance in our dataset, with a higher proportion of Paid versus Unpaid invoices, could contribute to a bias that favors predicting invoices as Paid, thereby affecting recall adversely.

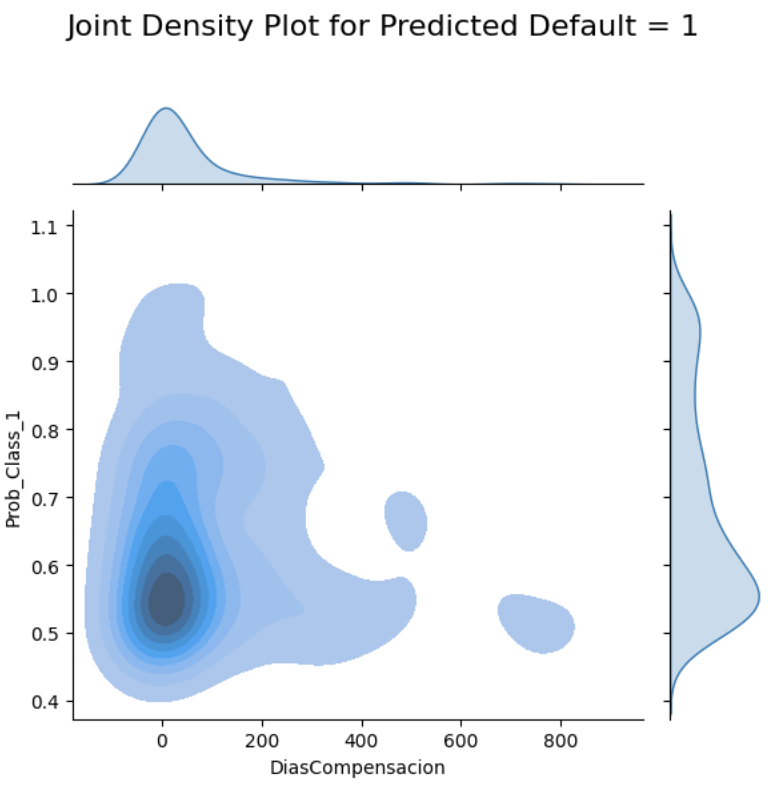
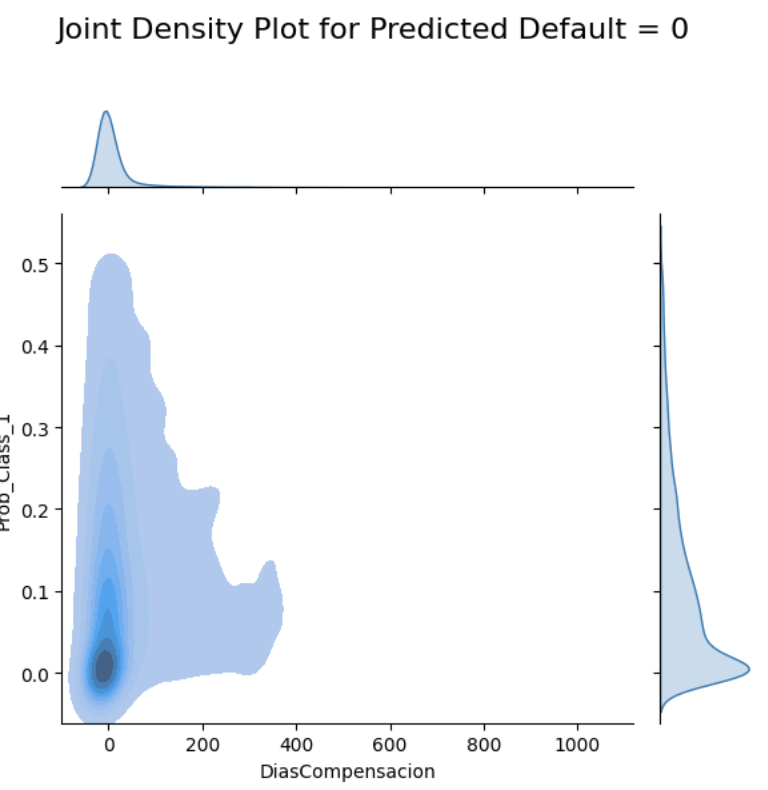
If the business requirement prioritizes a higher recall to more aggressively identify Unpaid invoices—even at the cost of a higher false positive rate—the optimal threshold based on the AUC-ROC curve analysis (for selected XGBoost) is 0.12. Adjusting the threshold to this level increases the recall to 80%, which means the model will correctly identify 80% of Unpaid invoices as such. However, this adjustment leads to a decrease in precision to approximately 27%, and the F1-score adjusts to 42%. While the accuracy and precision might appear promising, the lowered precision reflects the trade-off made to enhance recall.

****

This strategic choice can be useful in scenarios where the cost of missing an Unpaid invoice is significantly higher than the cost of wrongly classifying a Paid invoice as Unpaid. Such a strategy allows the business to manage risk more effectively, although it might increase the operational workload due to a higher number of transactions flagged for review.

We conducted an analysis by plotting DiasCompensacion (Day lag in payment from the invoice date) against the predicted probabilities for class 1, segmenting the data into two groups based on the predicted classes (0 for paid and 1 for unpaid) to verify if the model was learning in the expected direction. This analysis aimed to discern if there was a distinct pattern correlating the length of payment delay with the likelihood of default.

This variable, DiasCompensacion, was ultimately excluded from the model to prevent potential data leakage, as the days of compensation can only be determined post facto, i.e., once it is known whether an invoice has been paid. Ideally, we would expect a higher concentration of paid invoices associated with shorter compensation days and an increasing frequency of unpaid invoices as the lag extends. The analysis revealed trends closely aligned with these expectations, confirming the model's ability to learn and predict based on sensible economic and behavioral assumption



**4.4 Clustering**

We further segmented customers into groups based on key attributes such as individual or company status, credit scores, and vehicle ownership, which allows for the development of tailored models for each group. Our clustering analysis segmented 280,000 customers involved in 1 million transactions from 2021 to 2022 into 10 clusters, revealing significant heterogeneity in size and average default rates, which stood at 20.2% across the sample. Each cluster was analyzed to calculate the mean default rate and identify distinguishing features.

Cluster results are displayed in the appendix, tables 9-11. Significant findings include Cluster 0, which comprises 40% of the clients, showing a lower unpaid rate of 10%. Members of this cluster typically have a high Equifax score of 642, hold a single contract, engage primarily in toll operations, and reside in REGION 13, with a high average transaction amount of 2755. In contrast, Cluster 8, with an unpaid rate of 47%, includes customers with a lower average Equifax score of 571, two contracts, and similarly high transaction amounts but are involved in collection cost operations. In addition, Cluster 2 stands out with an unpaid rate of 18% and customers who are legal entities, have 1 or 2 contracts, and own an average of 18 vehicles. This cluster engages in either toll or interest operations and has a very high transaction amount of 3474.

1. **Overall Impact**

**5.1 Macroeconomic Analysis**

The primary objective of this analysis was to analyze how macroeconomic trends impact toll payment delays. By linking economic factors to payment behaviors, we aimed to provide crucial insights for stakeholders to anticipate shifts in default rates and support negotiations regarding toll pricing with the Chilean government. We used both univariate and multivariate regression models to identify significant correlations between macroeconomic indicators and payment delays. Our comprehensive analysis succeeded in revealing the impact that multiple economic variables (such as GDP, inflation, household debt, and urban transportation) had on default rates. This provided the sponsor with a robust economic rationale for anticipating payment defaults and initiating strategic toll rate renegotiations. The successful completion of a detailed macroeconomic correlation presentation, supported by visualizations and statistical evidence, represents a key milestone in this project.

Our findings offer immediate benefits to the sponsoring company’s operations and strategy. In the short term, the insights will allow the firm to anticipate future toll payment defaults with greater precision, enabling more proactive actions to minimize revenue loss. Additionally, the data-backed correlation analysis strengthens the firm’s negotiations with the Chilean government regarding toll pricing adjustments, directly supporting the sponsor's strategic goals.

The potential applications of this project extend beyond the current analysis. Future work could involve redefining the models by incorporating additional variables such as microeconomic indicators, thus capturing more nuanced relationships between economic factors and default rates. Additionally, the models could be adapted for use in other markets where similar toll collection systems are in place. By customizing our models for different economic environments, the company could apply our approach to create valuable insights and strategies across its global portfolio, ultimately improving the performance of its infrastructure investments. Furthermore, continuous monitoring and updates to the models will enable the sponsor to adapt to emerging economic trends, fostering a data-driven approach to decision-making that ensures a long-term competitive advantage.

**5.2 Predictive Model**

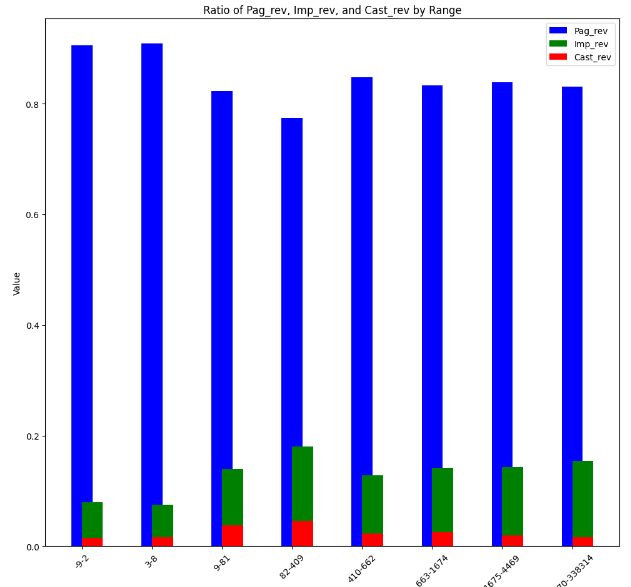
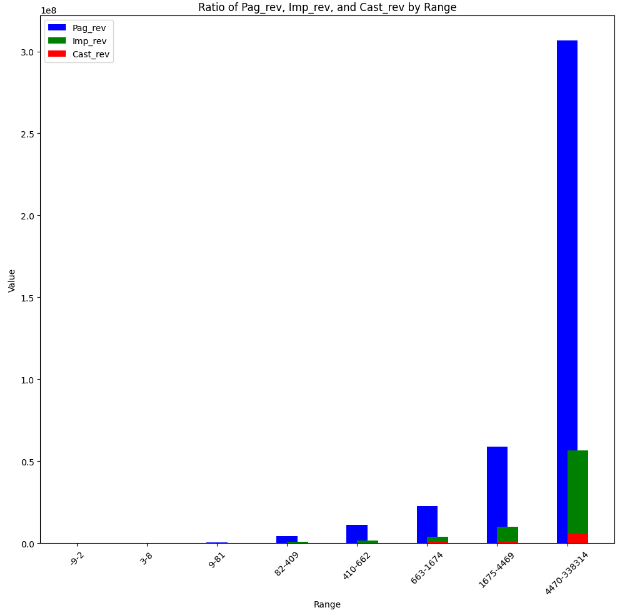
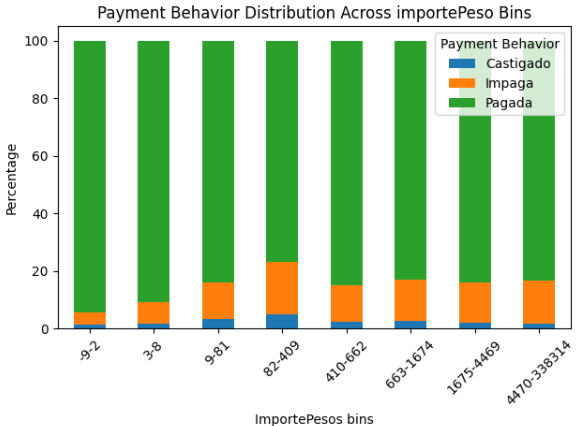
The objective was to develop a predictive model that could alert Vespucio to potential risks during the collection process. Not only did we successfully meet this objective, but our analysis also provided deeper insights into the factors influencing invoice defaults. These insights will assist Vespucio in making informed decisions before proactive collection, ultimately reducing potential risks. To improve prediction performance, we have few recommendations as next steps:

* **Enhanced City Information:** Most models identified city-related variables as significant predictors of default. By integrating macroeconomic factors (as analyzed in the project) specific to these cities into our model as new explanatory variables, we could refine our predictions and better understand regional economic influences on default rates.
* **Inclusion of Customer Demographics:** Adding demographic variables such as age, gender, marital status, and salaries could enrich the model's data landscape. Care will be taken to ensure these features are incorporated in a way that minimizes bias and maintains fairness, potentially enhancing the model's predictive accuracy.
* **Advanced Feature Engineering**: Developing new features that encapsulate historical payment behaviors could provide a more nuanced view of risk at the invoice level, thus allowing us to capture patterns that might not be evident from raw data alone.
* **Segment-Specific Models**: Given the customer segmentation into distinct clusters, designing and training separate models for each cluster could lead to more precise default predictions by tailoring the model's focus to the specific characteristics of each segment.
* **Time Series Analysis:** Employing time series analysis models, such as ARIMA, might uncover temporal patterns and seasonality in defaults, which could enhance the predictive power of our models by integrating these temporal dynamics.
* **Model Explainability:** Tools like SHAP and LIME have been instrumental in revealing the influence of each feature on our predictions, helping us understand the contribution and direction of each feature towards the prediction, thus ensuring that our model's outputs are interpretable and aligned with business knowledge. Continuing to validate these findings through business-sense checks will confirm the practical reliability of our models.

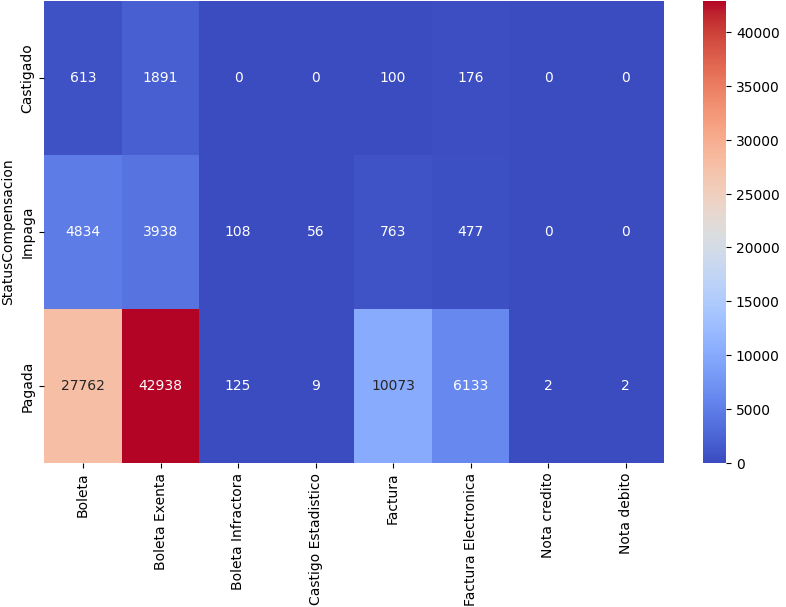
# 

# **Appendix**

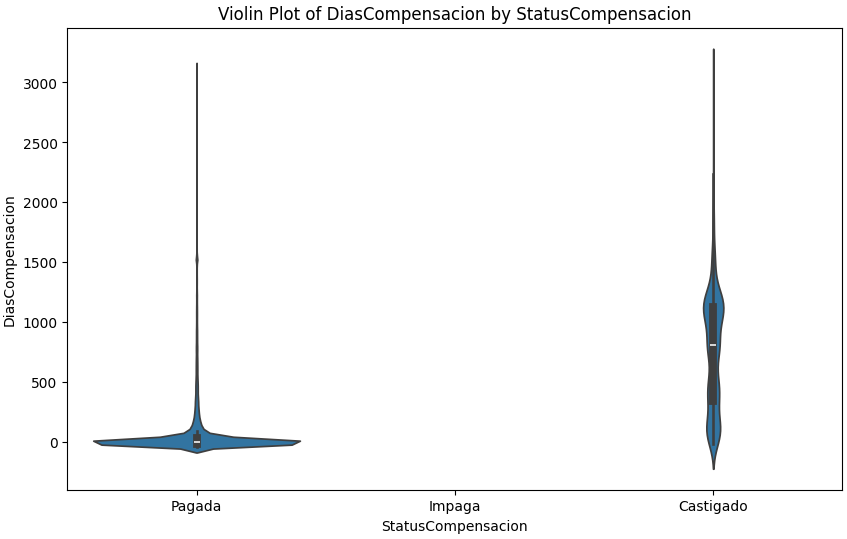
Graph 1-3: EDA: Transaction Amount (ImportePesos) & Payment Status



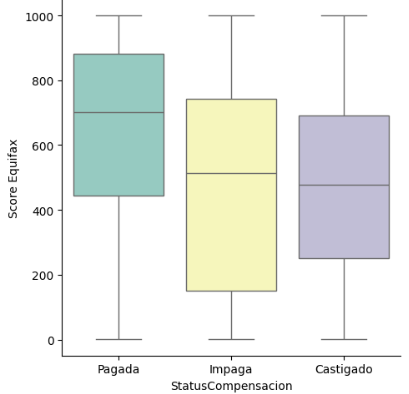
Graph 4: EDA: Document Type & Payment Status



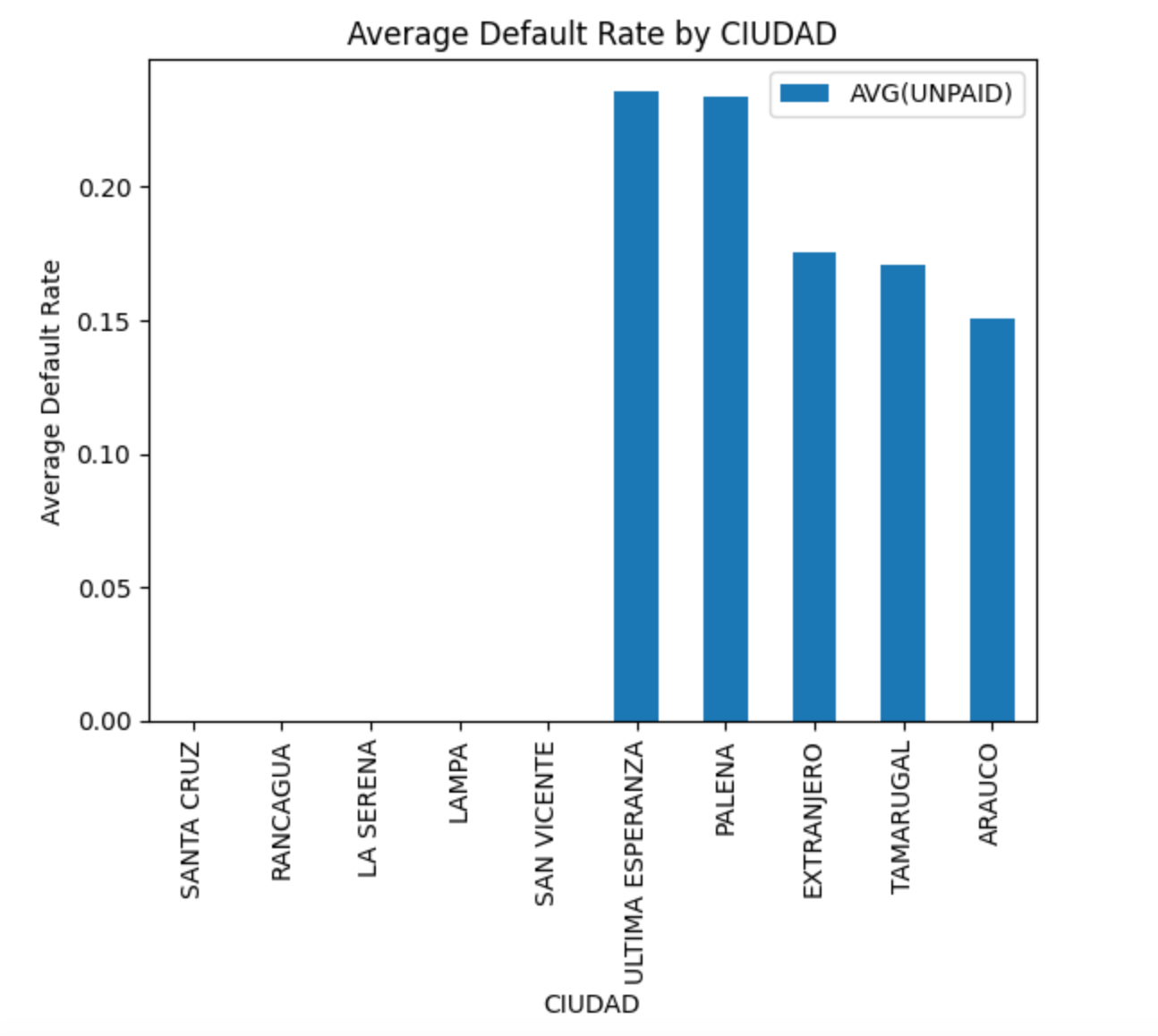
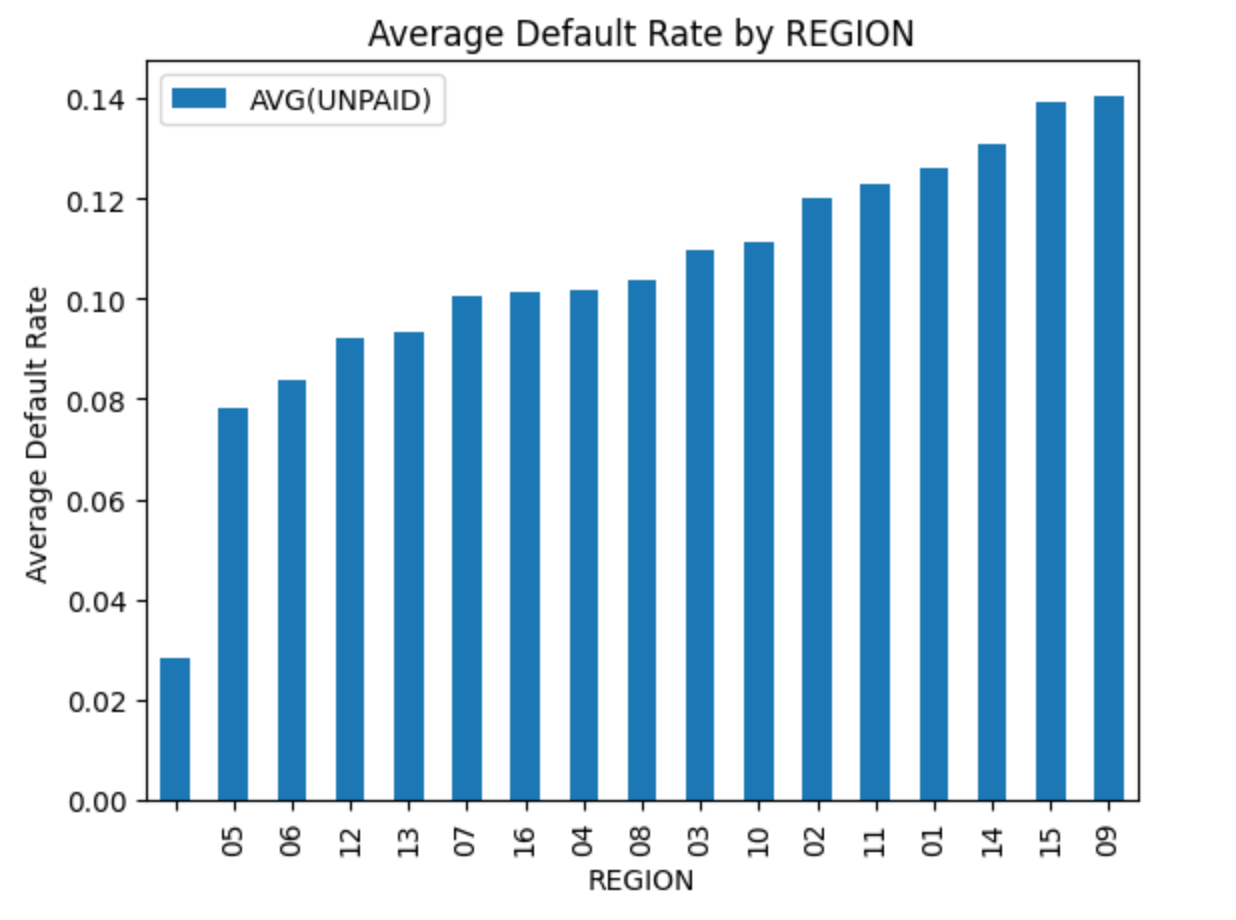
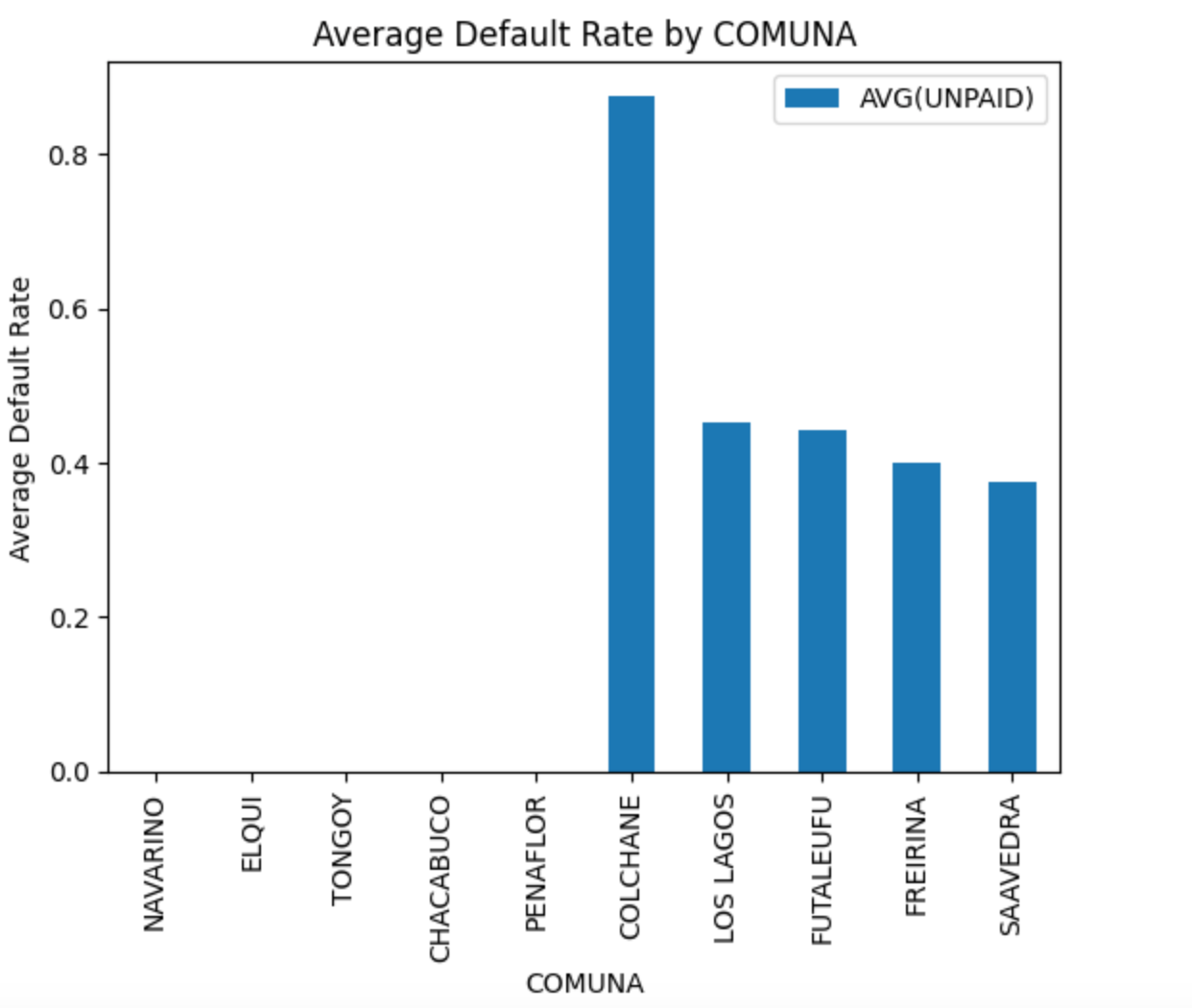
Graph 5: EDA: Payment Delay & Payment Status



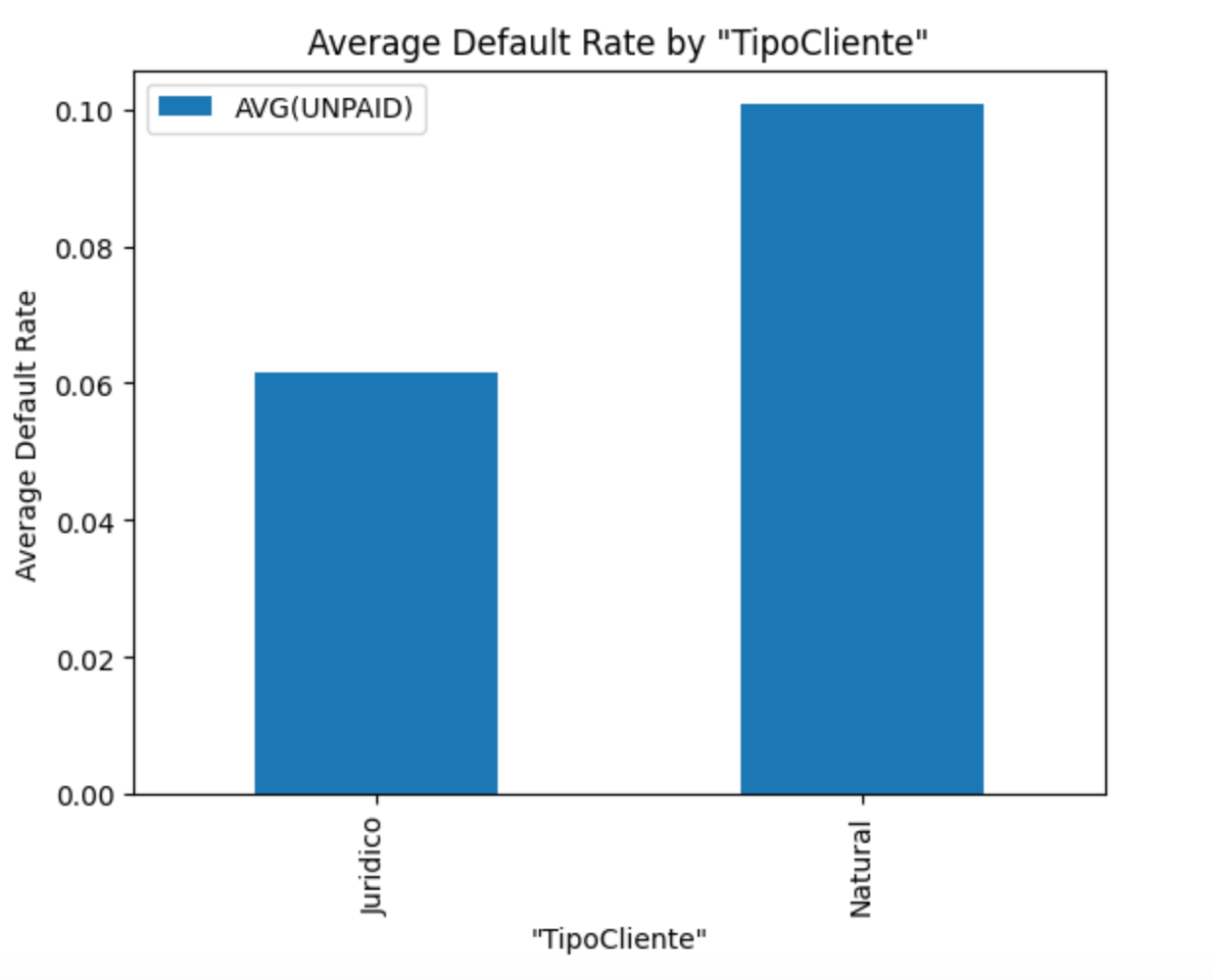
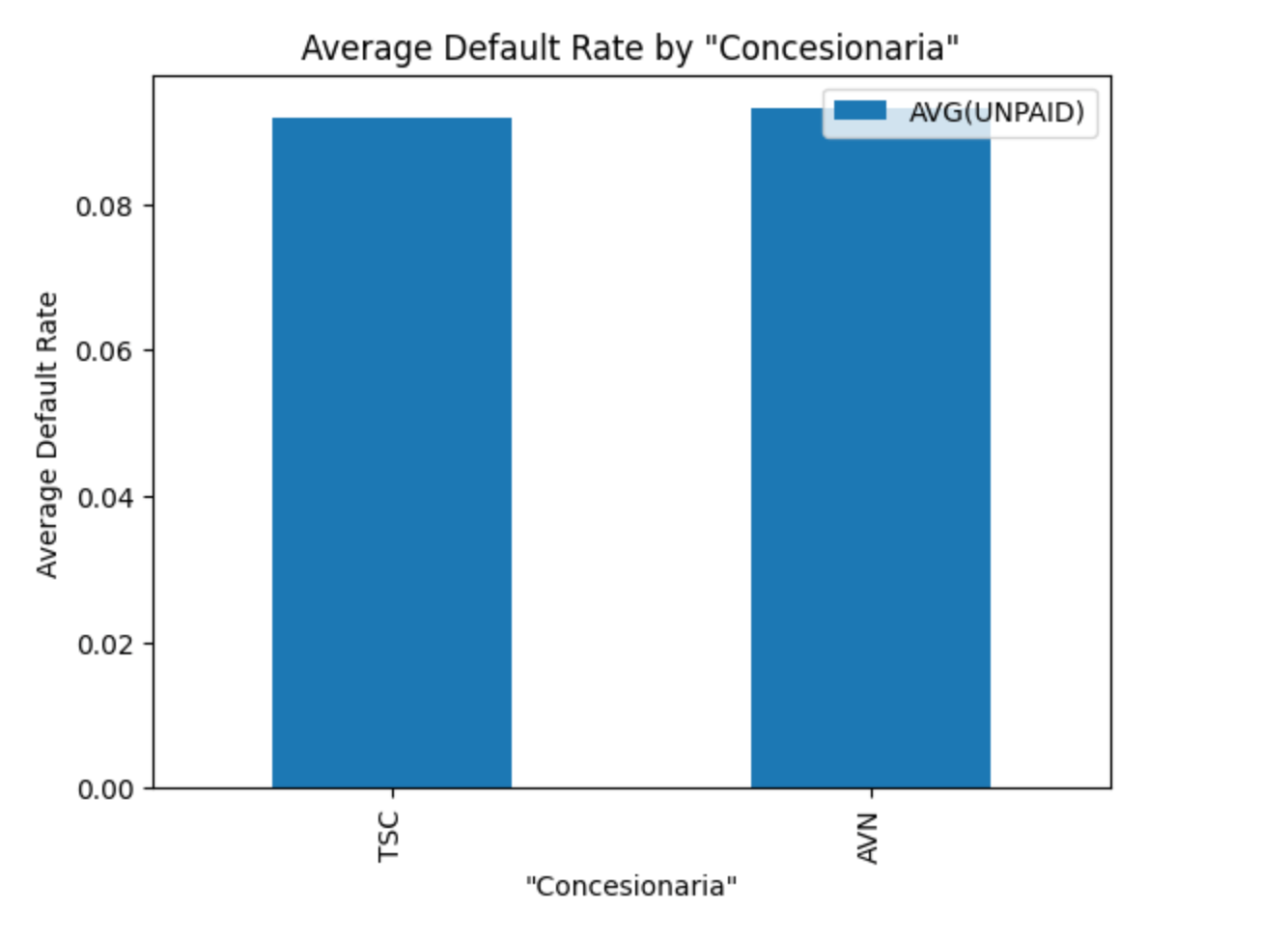
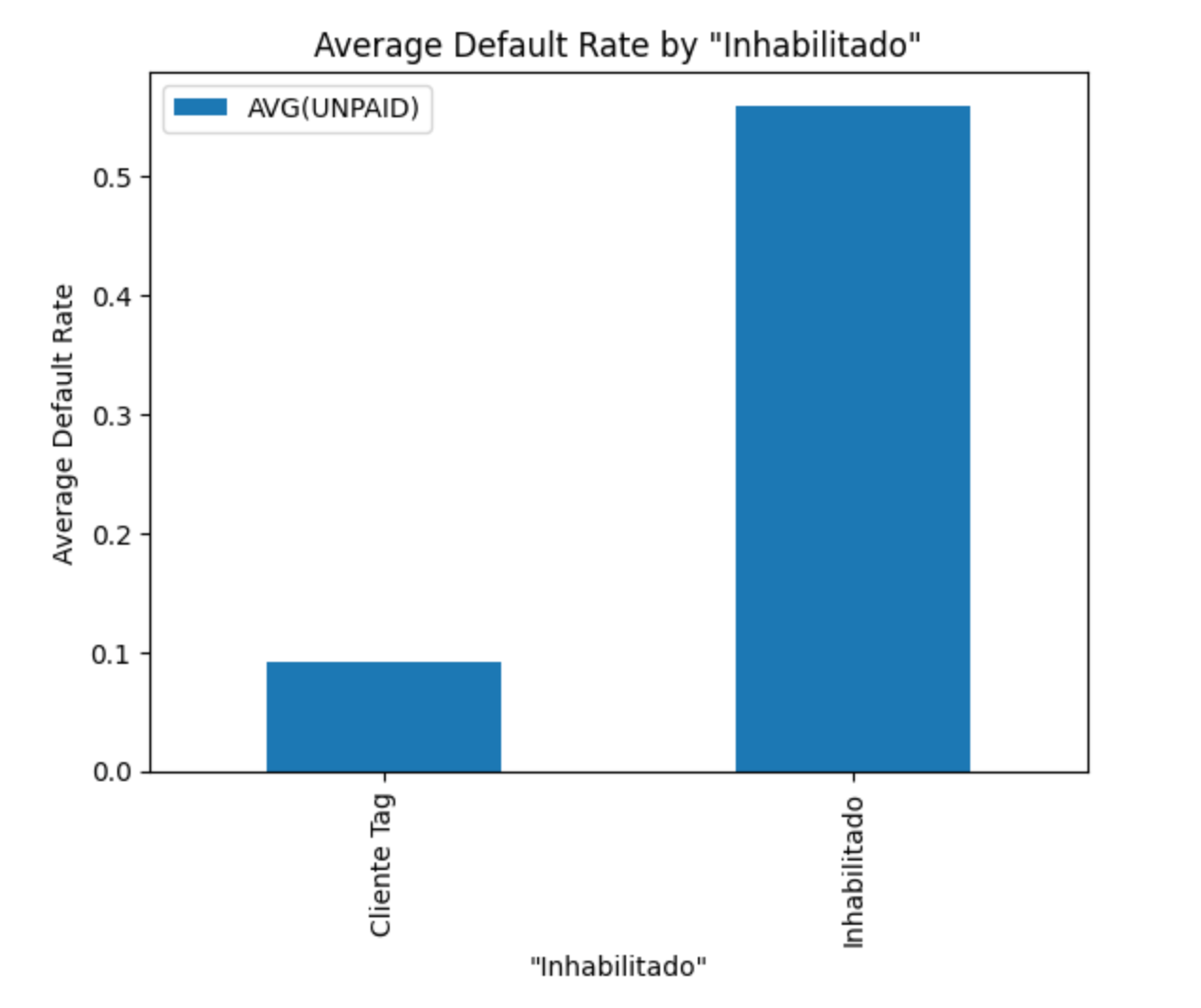
Graph 6:EDA: Equifax Score & Payment Status



Graph 7-9. Clustering Feature Selection: Geographical Variables



Graph 10-12. Clustering Feature Selection: Individual-Level Variables



Graph 13-15. Clustering Feature Selection: Continuous Variables

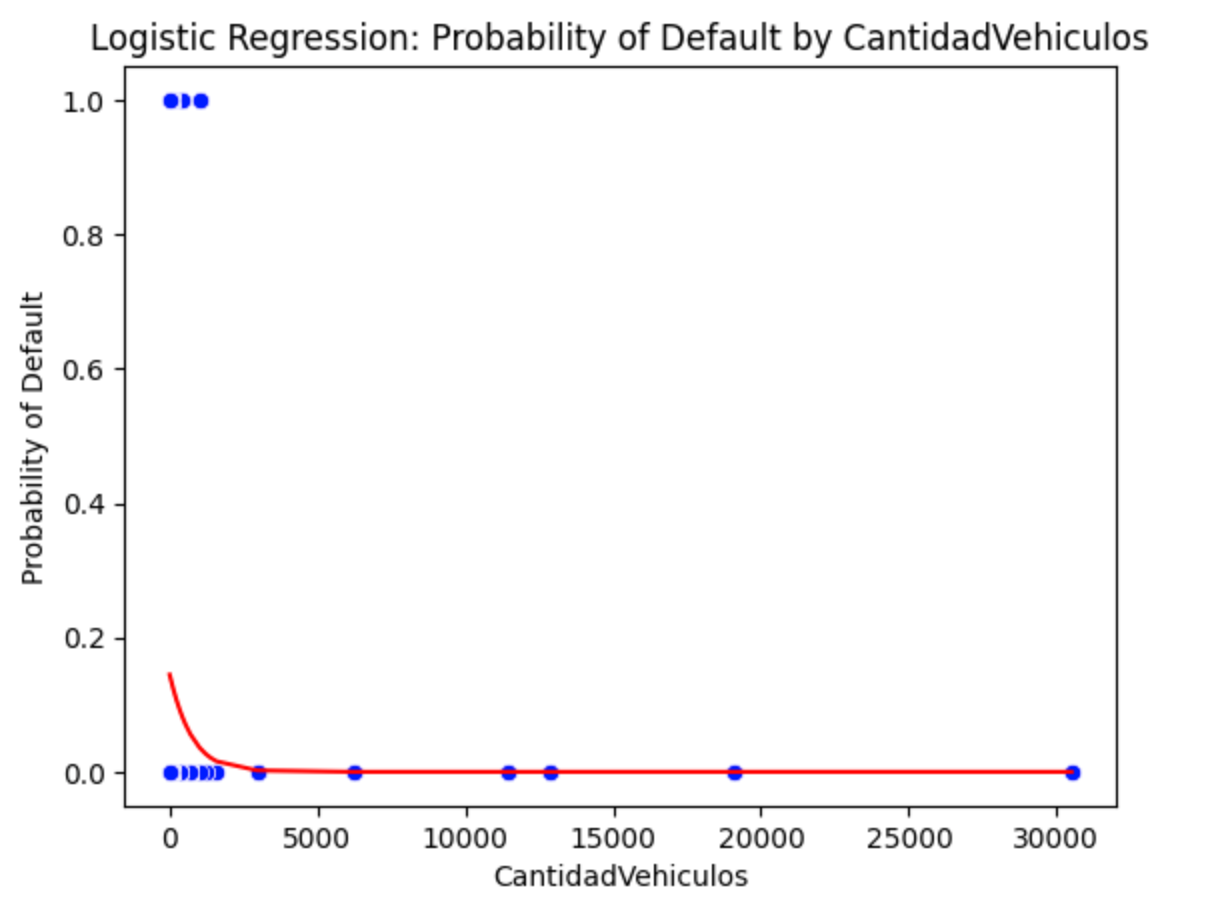
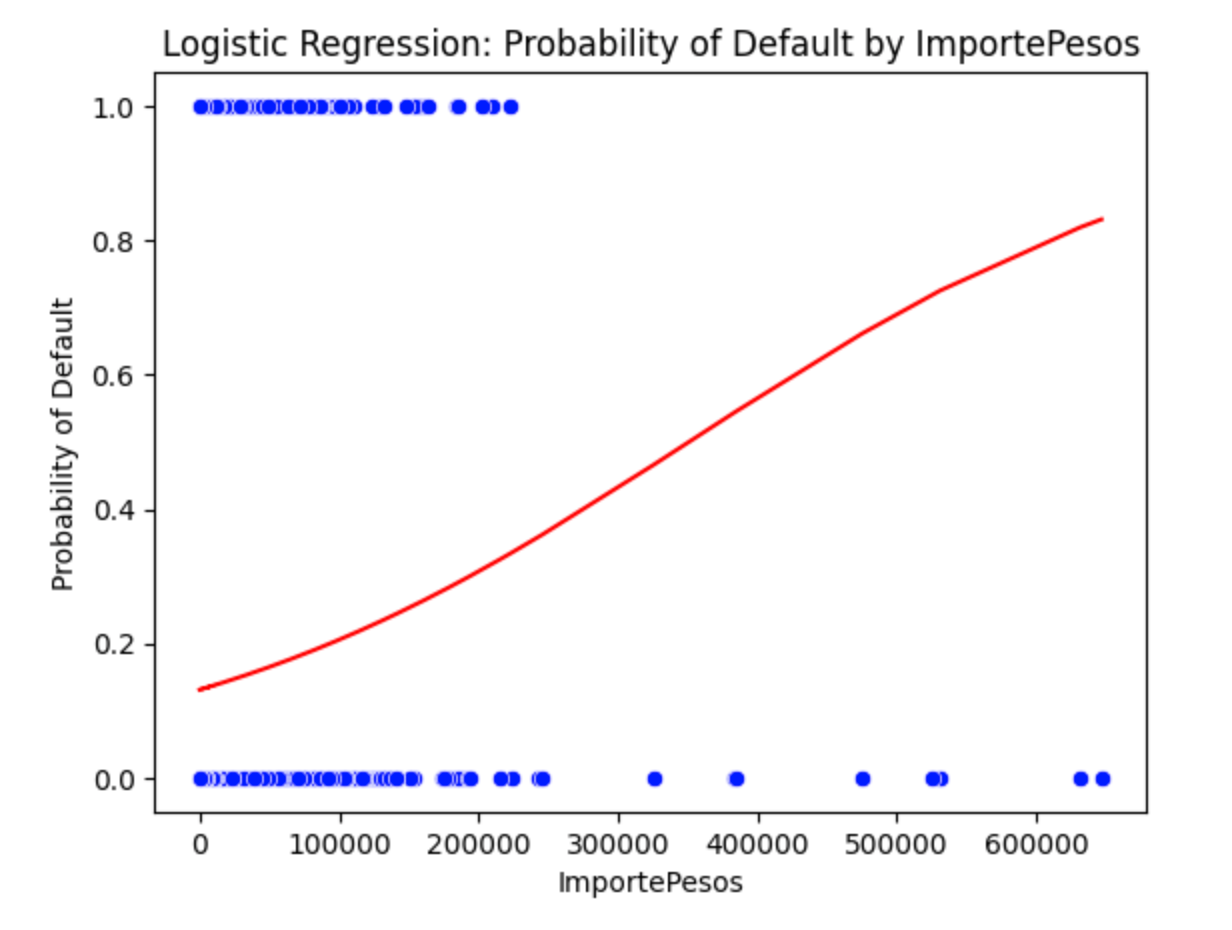
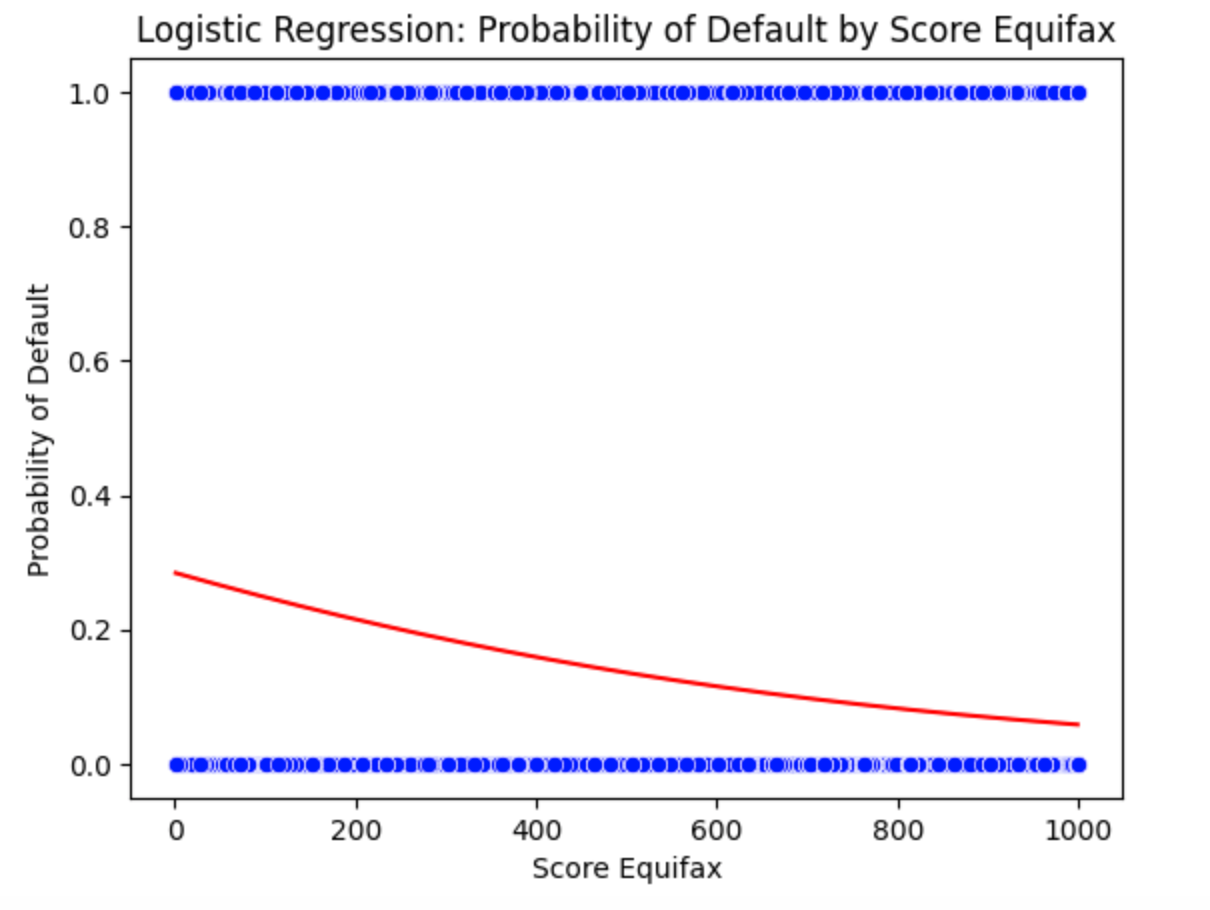


Table 1. EDA: Individual-Level Payment Status History

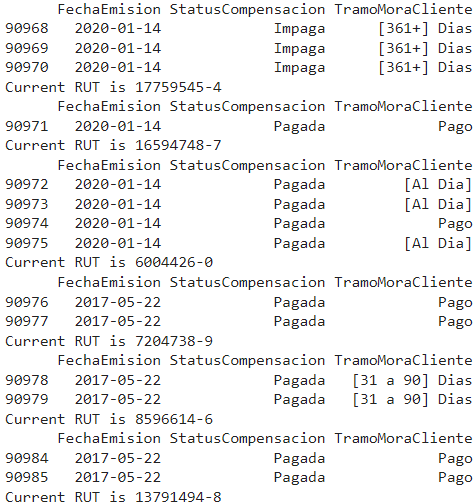


Table 2. EDA: Document Type & Payment Status



Table 3-4. EDA: Product Type & Payment Status

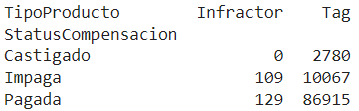


Table 5. EDA: Payment Type & Payment Status

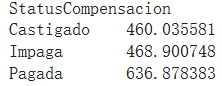
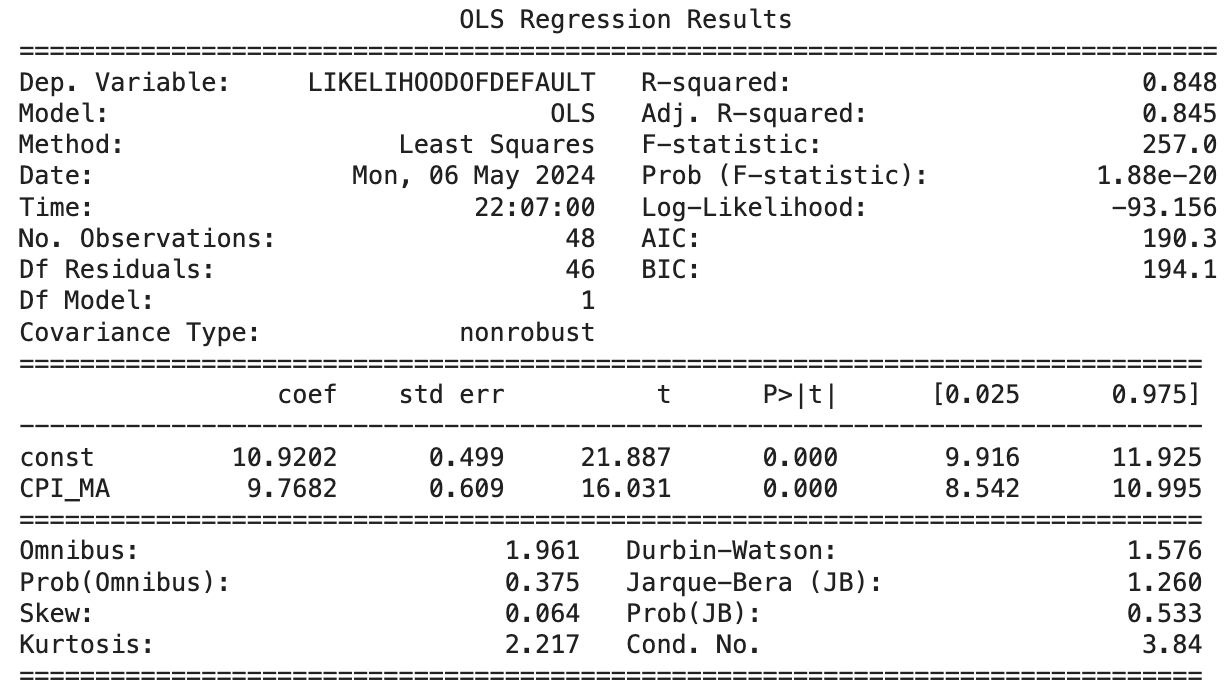
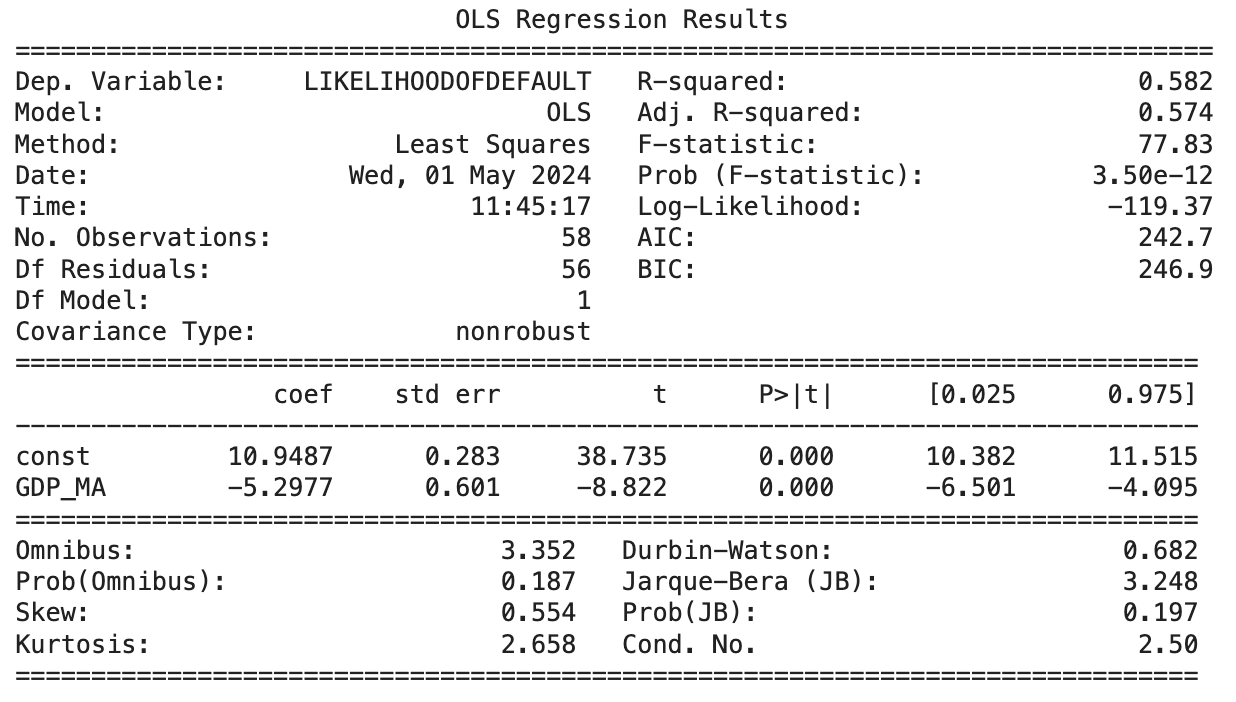


Table 6-9. Univariate Macroeconomic Model Results



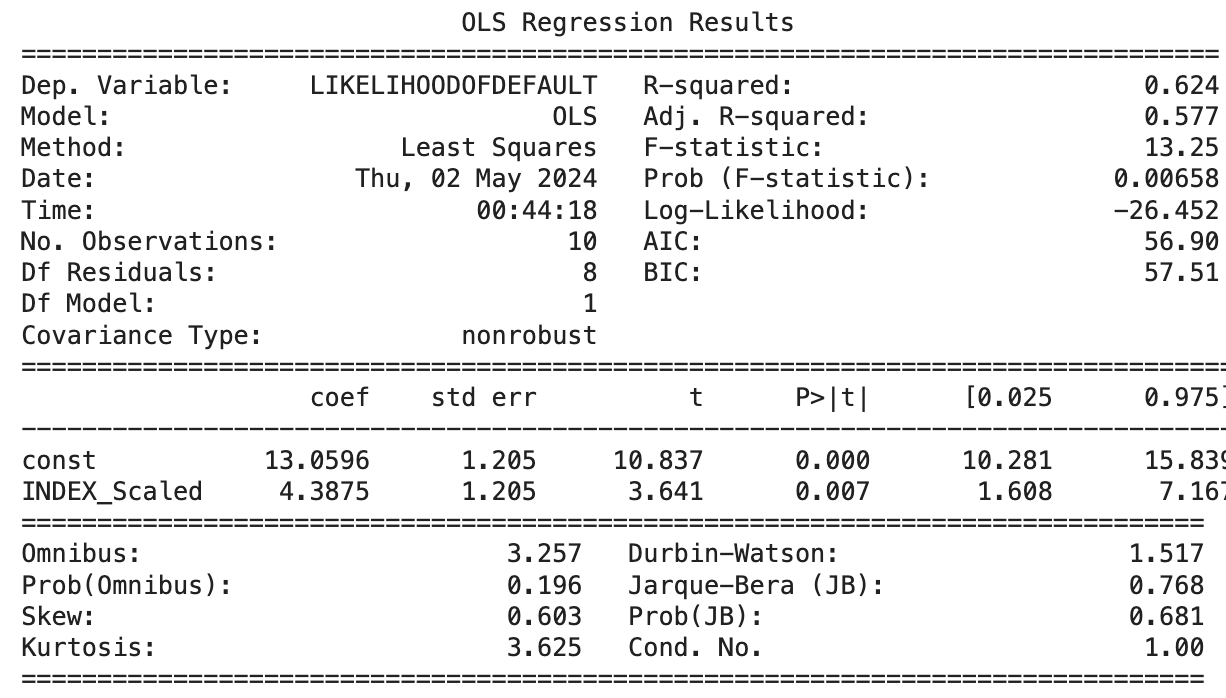
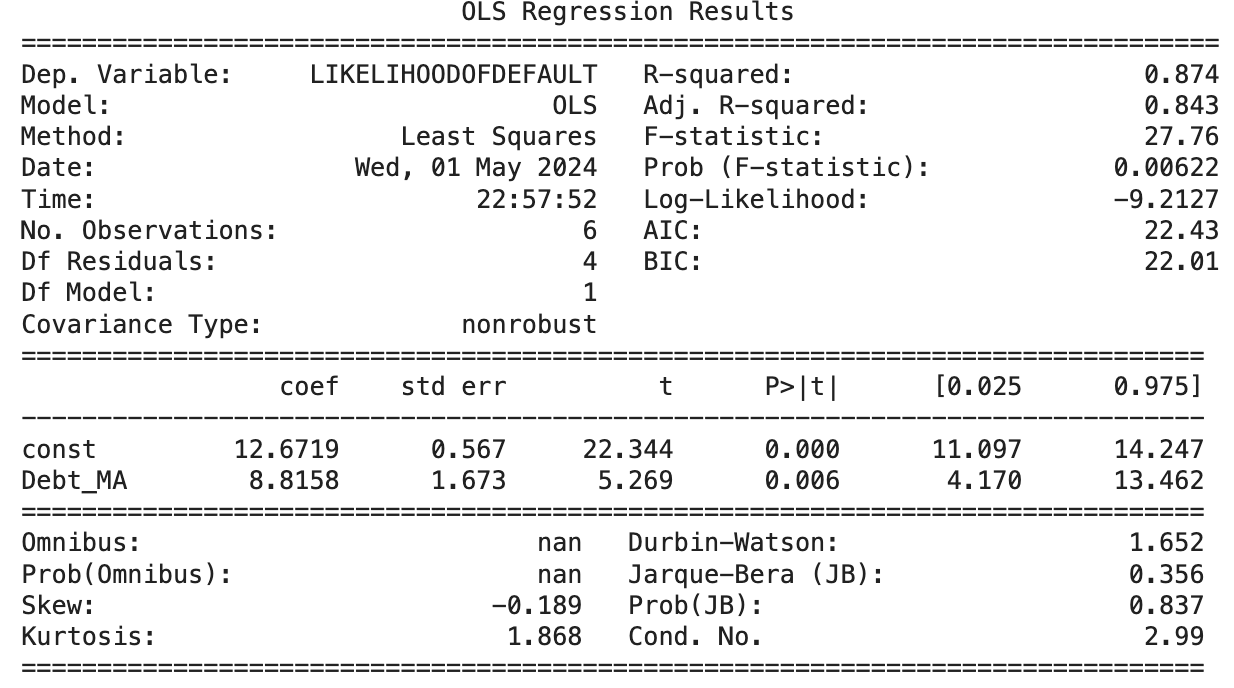
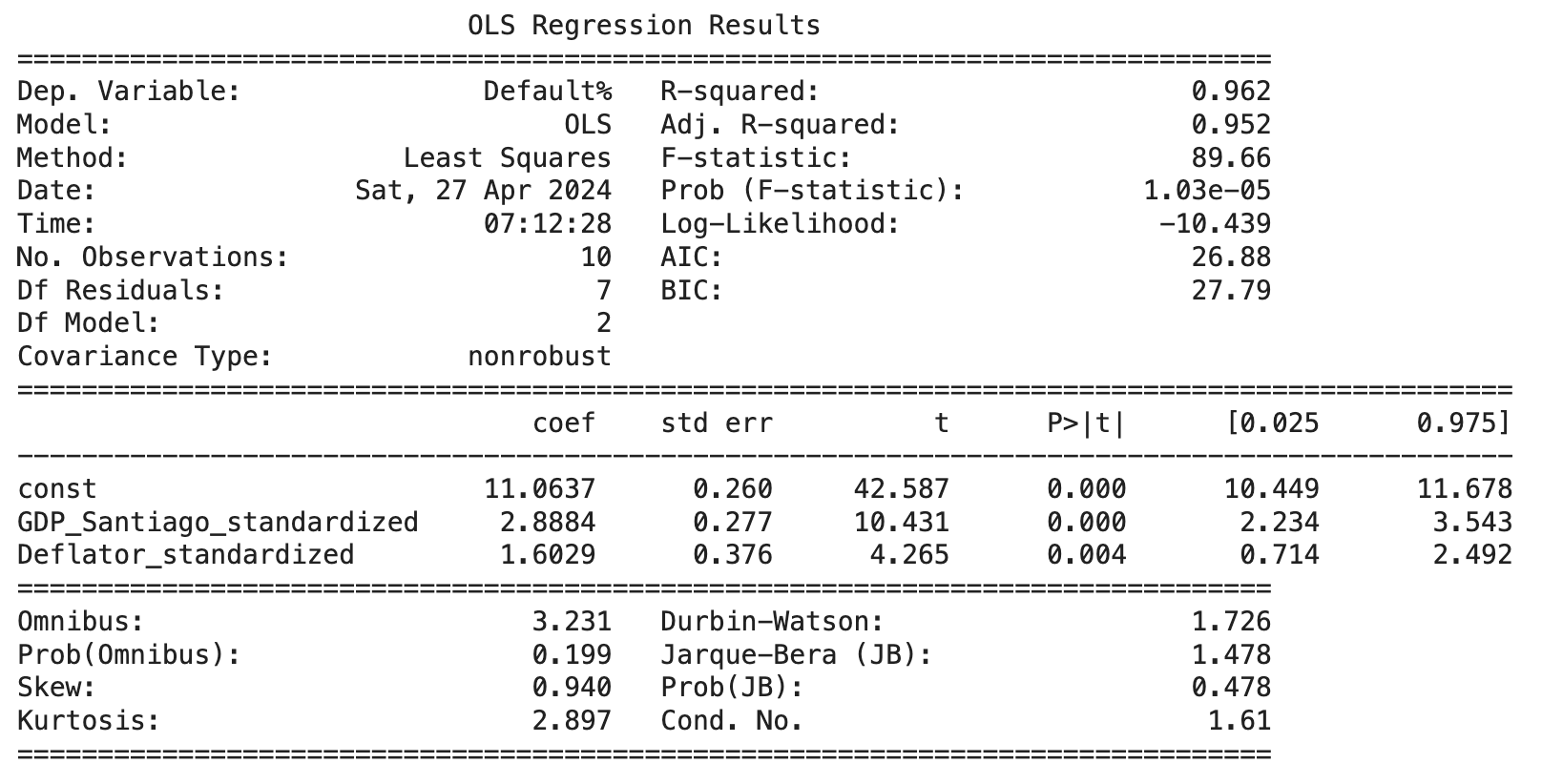
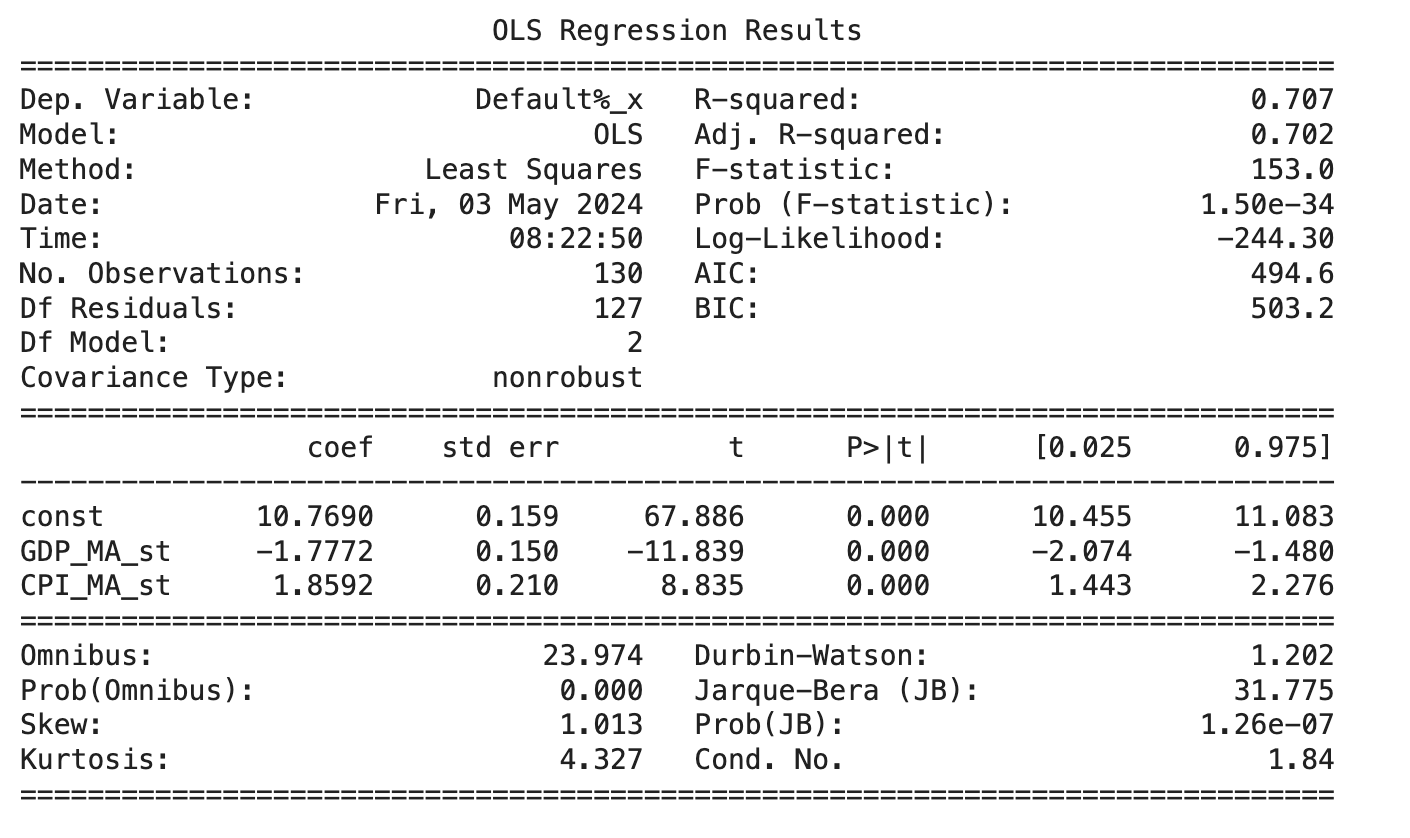
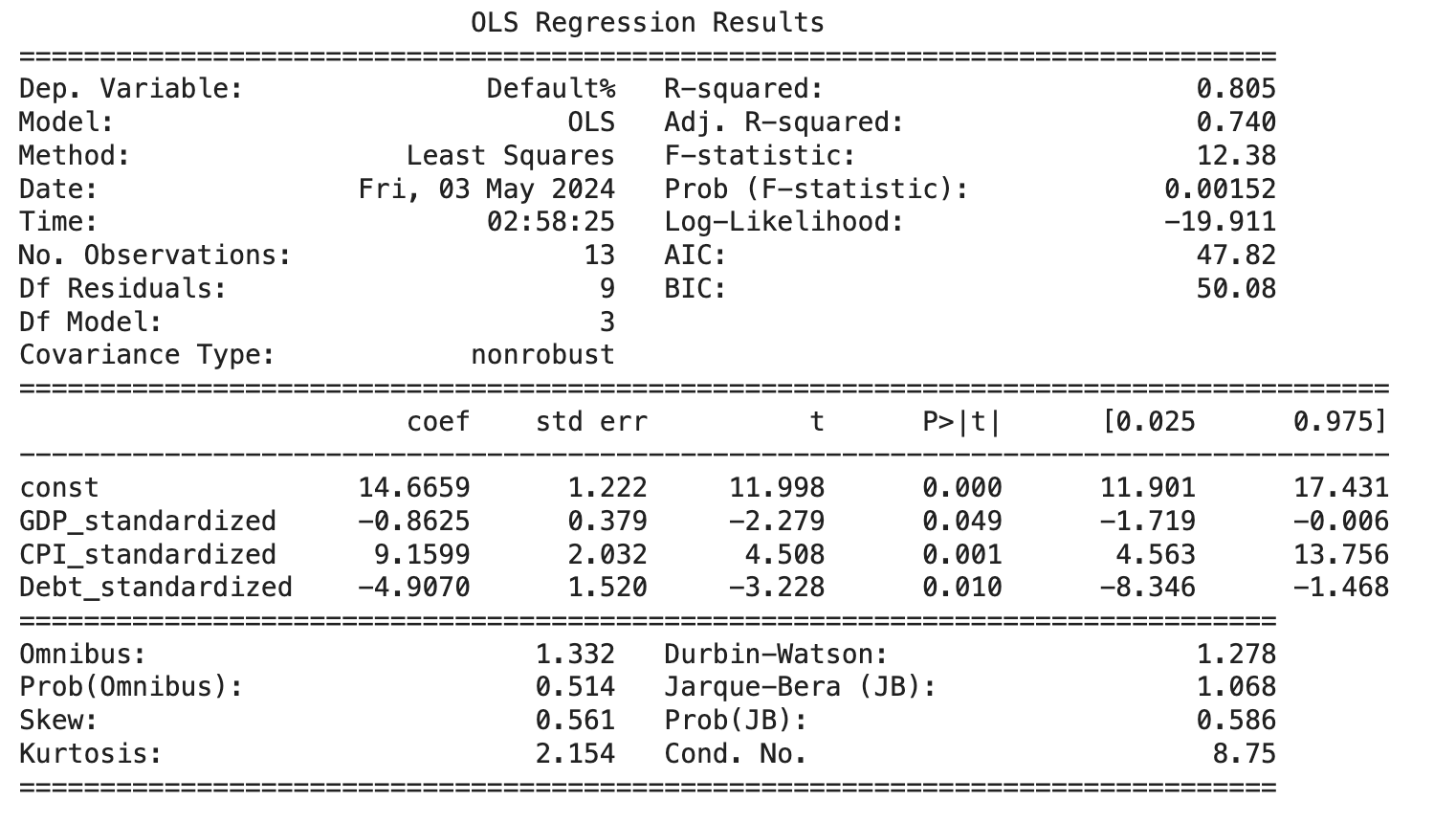
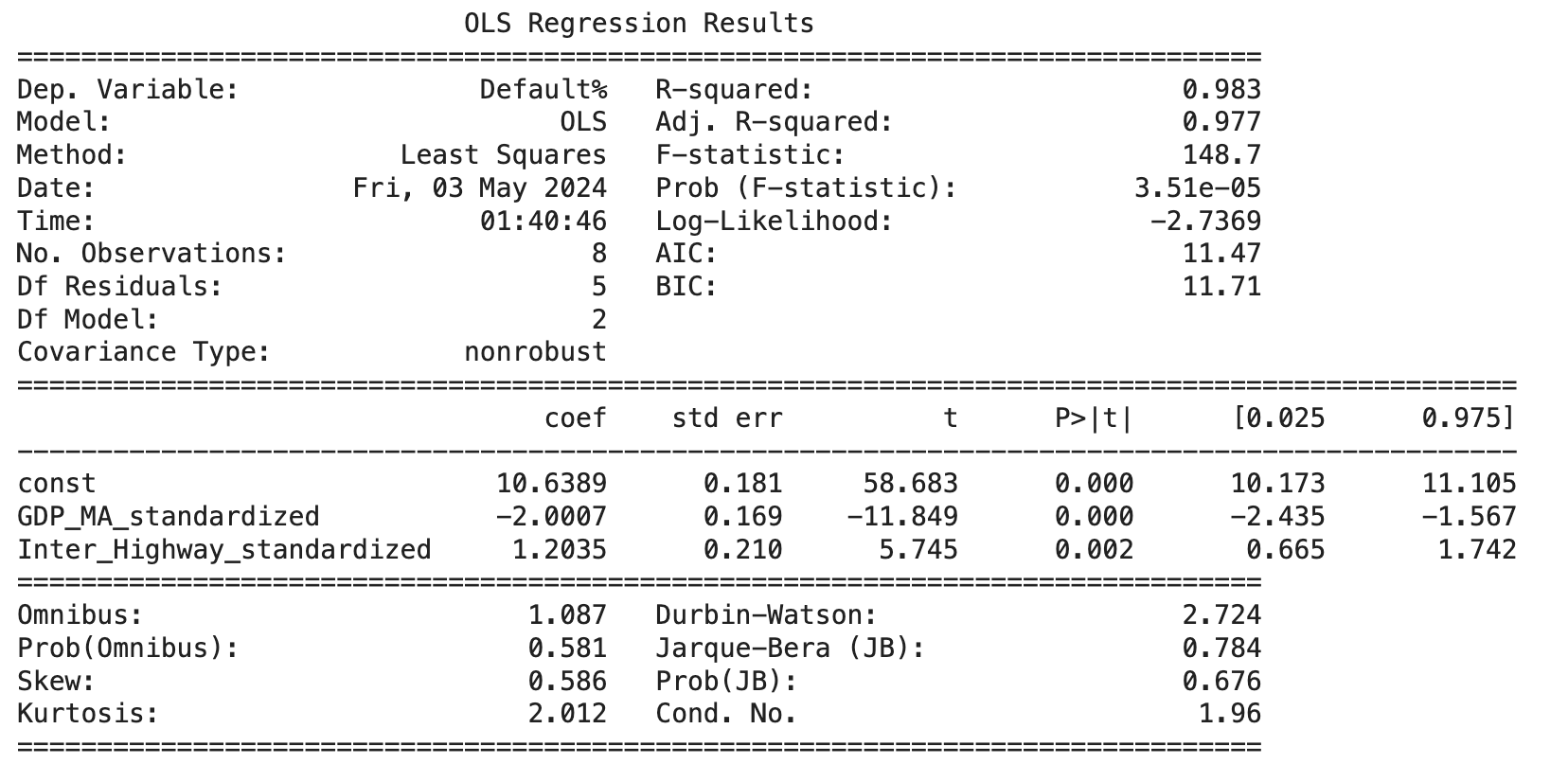


Table 10-13: Multivariate Macroeconomic Model Results





Tables 9-11: Clustering Results

