## 25752 Bank Lending and Analytics

### **1. Data-preprocessing**

#### A. The role of borrower income on credit risk prediction

The borrower’s income is very important in credit risk prediction. First of all, it is a key factor in assessing a borrower’s ability to repay loans. In general lending principles CAMPARI, borrower income is a part of the "Ability" criterion (University of Technology Sydney, 2025). Banks rely on the borrower’s income to ensure that the borrower has adequate cash flow to meet debt obligations. Secondly, bank lenders can evaluate a borrower’s financial capacity through their income, and its stability over time is essential in determining creditworthiness (University of Technology Sydney, 2025). In evaluating a personal credit, the central task is to ascertain the borrower’s capacity to repay. High, stable income levels reduce the likelihood of default, as they increase the probability of consistent repayment. Conversely, low or unstable income indicates a higher risk of default. Credit scoring models often incorporate income alongside other financial data to estimate the repayment capability (University of Technology Sydney, 2025). Moreover, in commercial banks, the logistic regression model effectively predicts digital loan defaults with income to loan ratio and credit score being critical variables (Barasa et al., 2025).

#### B. Borrower incomes by state in the US

### **Import data**

The data of Median Household Income by State: 1984 to 2023 is divided into 2 dataset by Current dollars (current.csv) and 2023 dollars (2023.csv). Firstly, consider the dataset of current dollars.

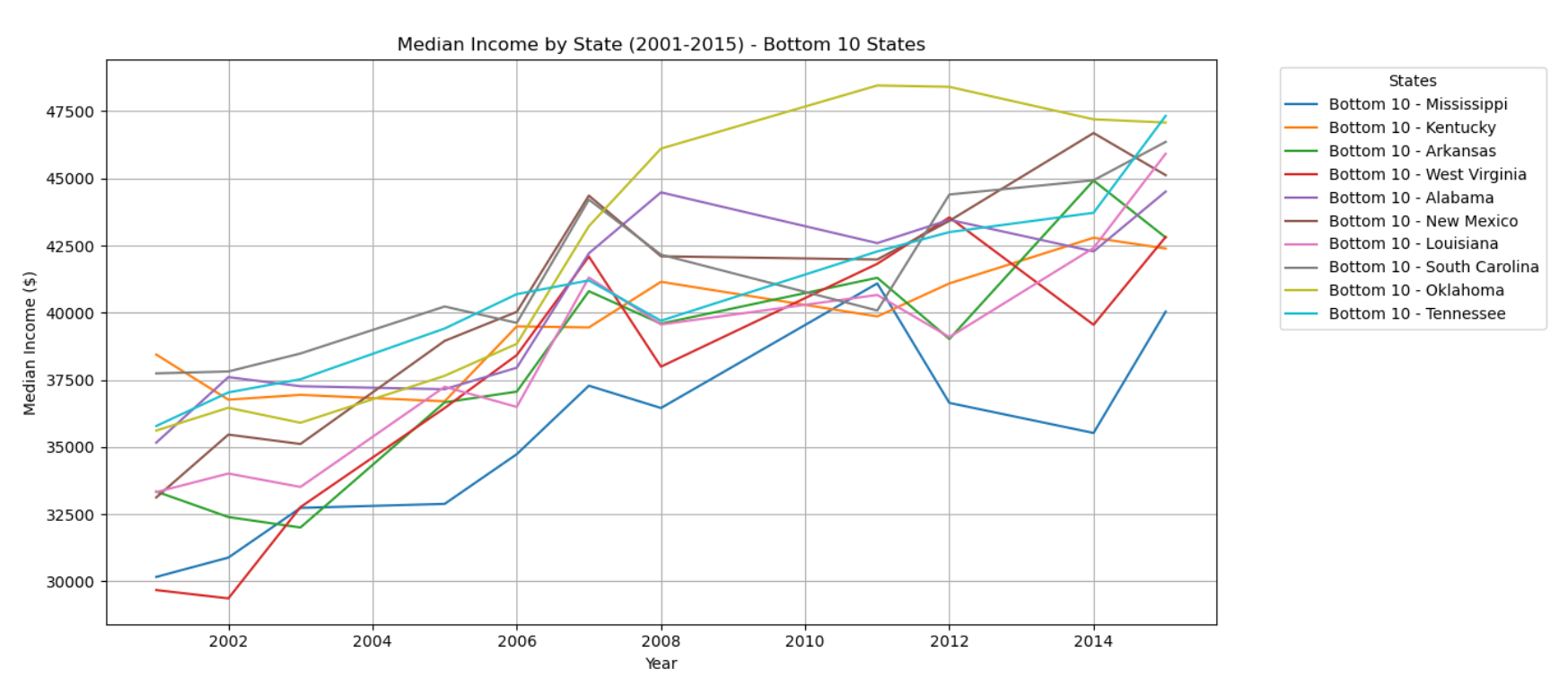
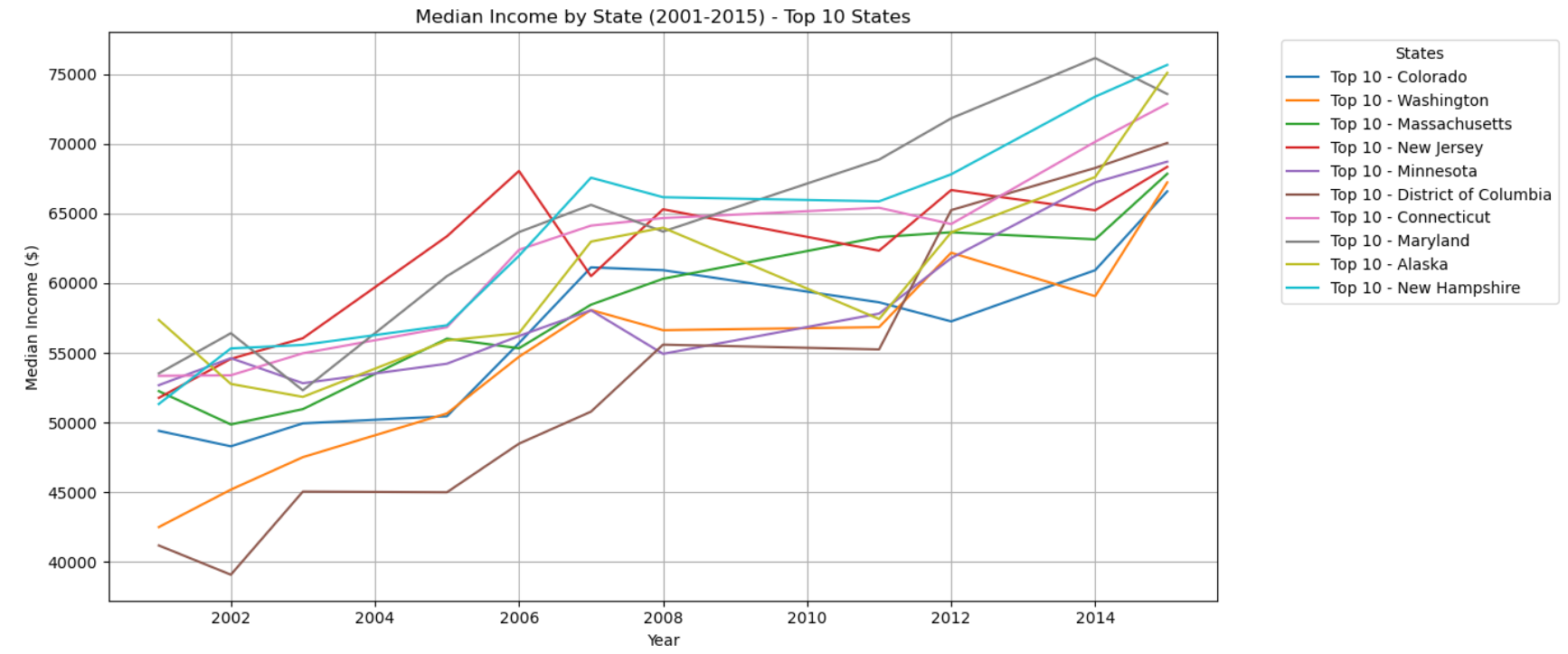
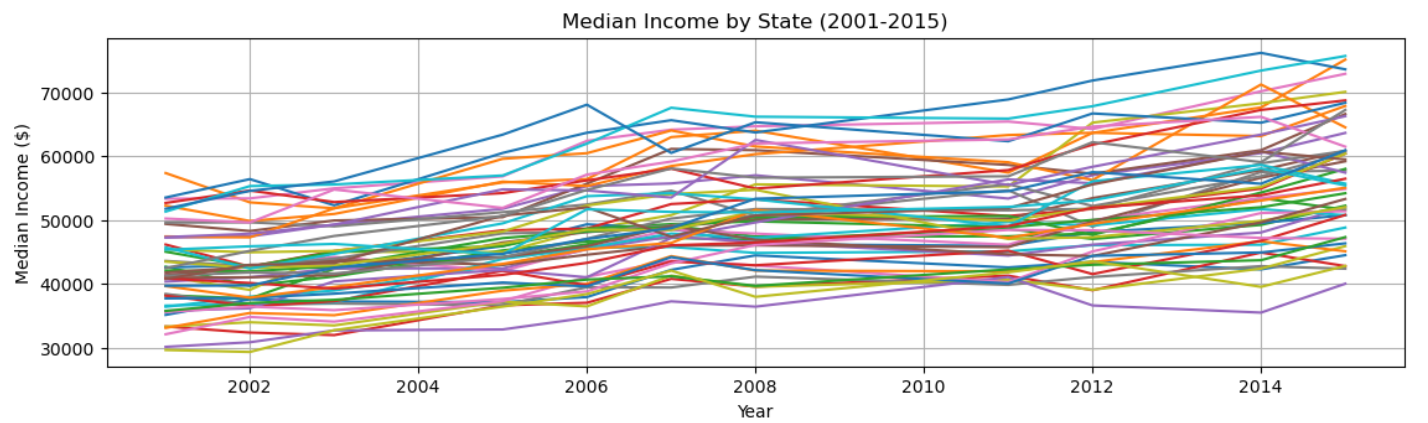
**Create a new dataset including 4 columns: State, Year, Median income and Standard Error**

### **Collect time series from 2001 to 2015 of borrower incomes by state**

### **Describe the data by matplotlib**

### **Plot time series by states**

When plotting data for all states, the graph becomes overcrowded and difficult to interpret, so plotting the Top 10 and Bottom 10 states simplifies the data and allows us to focus on the most meaningful trends.



### **Interpret outputs**

Overall, there is a general upward movement in median income over the years. However, the rate of increase is different from state to state. In the top 10 states, states such as Colorado, Washington, and Massachusetts show strong upward trajectories over the period, with median incomes increasing steadily over the years. This suggests economic growth in these states and positive impacts from some factors, such as employment opportunities, stronger industries, etc., may push wages up. Moreover, these states could have benefited from policies or economic conditions, which also suggests a well-educated, high-skill labor force with higher wages.

On the other hand, for the bottom 10 states, they show sluggish growth. This indicates slower economic growth or stagnation, which could be due to factors such as weaker industries, lower educational attainment, or economic challenges. Mississippi, in particular, stands out with one of the lowest median incomes throughout the period, which may reflect persistent poverty or challenges in the local economy.

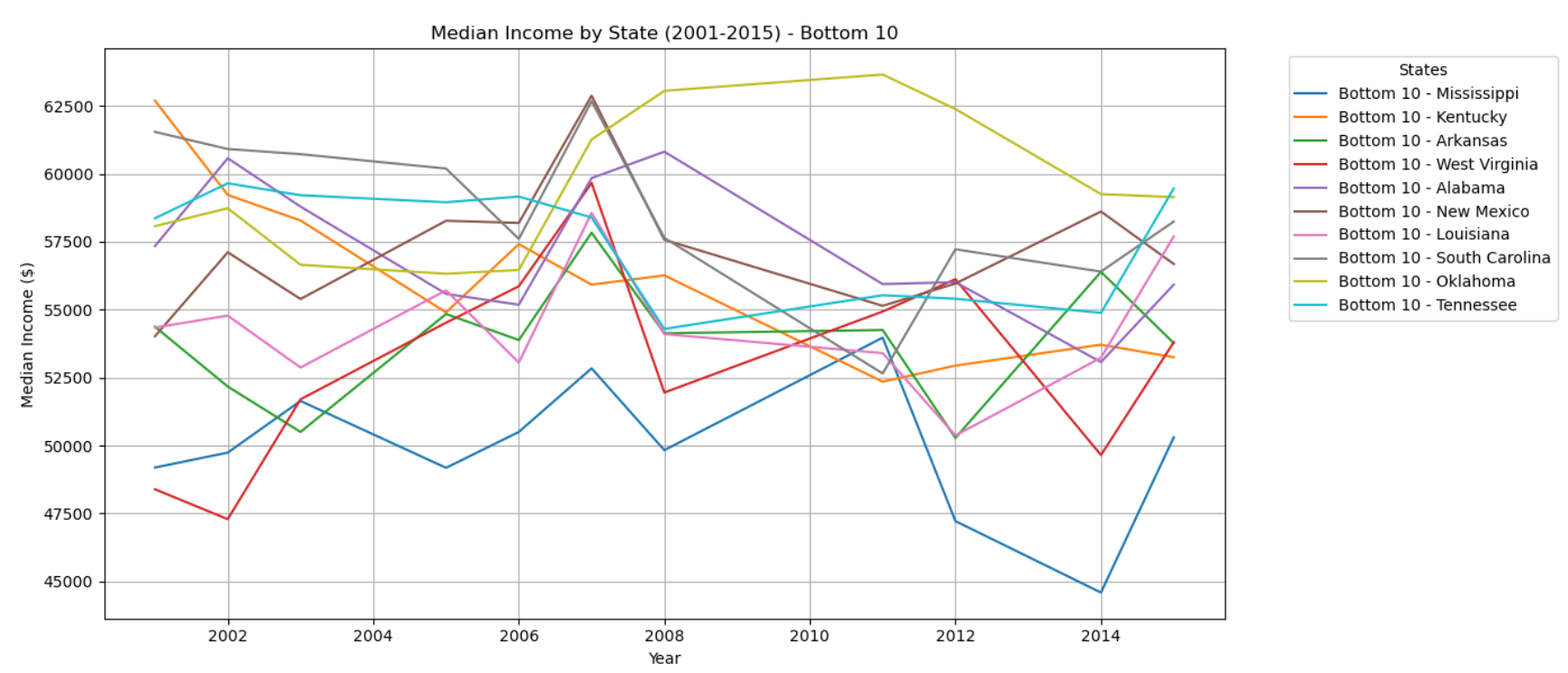
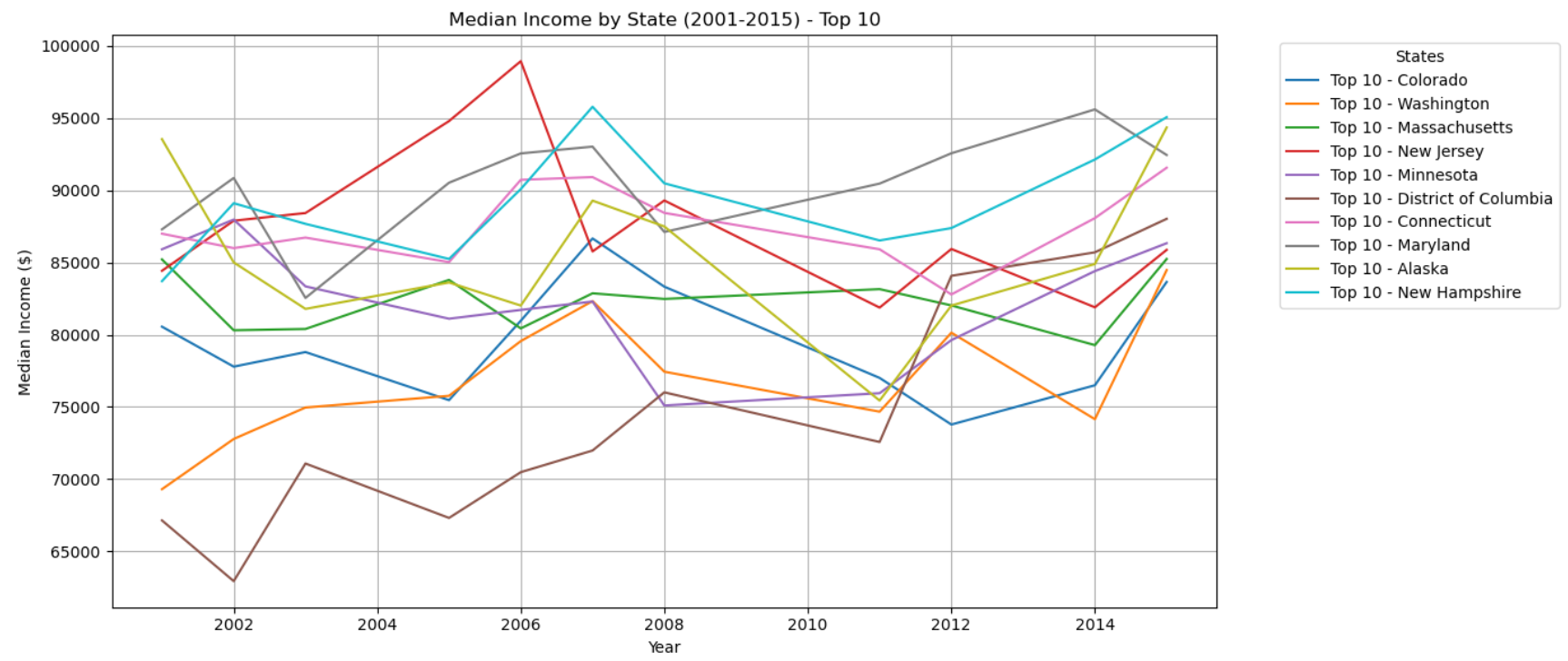
The significant gap between the top and bottom states highlights a growing regional income disparity. Economic strategies focusing on industry diversification, education, and higher wages could help narrow the gap between these regions.

### **Interpolate missing years**

Because the data shows a consistent pattern or trend over time, linear interpolation is used first to create an estimate of the missing data based on surrounding trends. Next, Forward-fill is used afterward to fill any gaps that still exist, ensuring that data continuity is maintained as well as no missing values persist for those years.

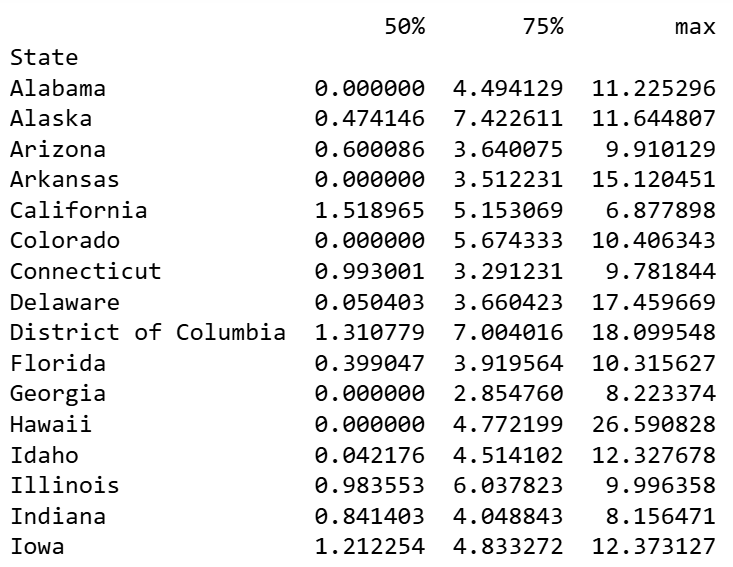
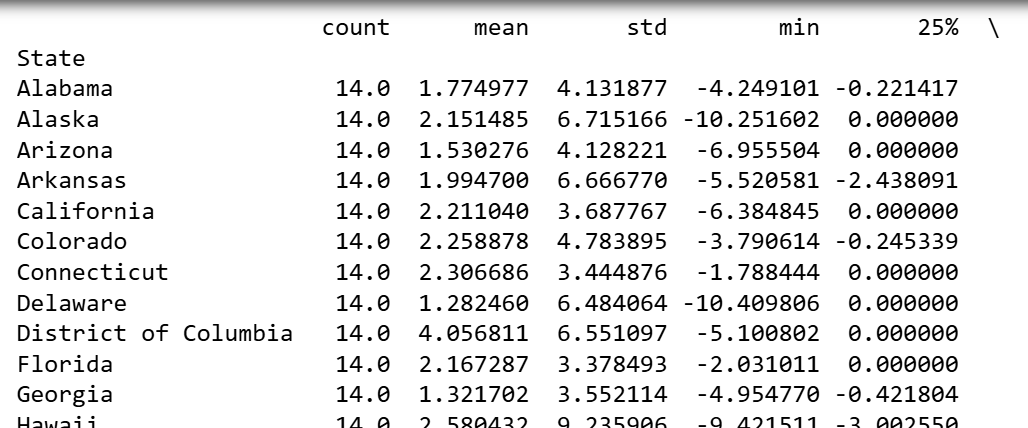
### **Import and interpolate 2023 data**

Similar to current dollars, it can be seen that in 2023 Dollar most states show an upward trend in median income. Some states have seen sharp increases in median income, while others show modest gains or stagnation. However, in the bottom 10, unlike the top states, which show clear and consistent growth, the bottom 10 states have relatively flat or declining lines. This suggests that these states have struggled with economic growth in terms of median income, even after adjusting for inflation. In particular, some states show sharp drops in median income around the 2008 financial crisis (e.g., Mississippi, Arkansas). In general, this reflects growing income inequality across the U.S. By adjusting for inflation to 2023 dollars, it's clear that the nominal income in bottom states might have increased, but the actual purchasing power has remained stagnant or grown slowly, especially when compared to high-income states.



#### C. Growth rate of income

### **Describe the heterogeneity (differences) in growth rates**



The mean growth rate for each state provides an indication of the average income growth over the period. We can see that Iowa has the highest average growth rate of 2.93%, while Kentucky has the lowest average growth rate of 0.75%. In addition, std (standard deviation) measures the variability of income growth across years for each state. For example, Hawaii has a high standard deviation of 9.24%, indicating that income growth in Hawaii has been highly volatile, with large fluctuations between years. Some states like Wisconsin have very low variability (1.56%), suggesting that income growth has been relatively steady. Next, percentile distributions (25%, 50%, 75%) show the range of growth rates. For instance, in 25 percent, Georgia with -0.42% suggests that a significant portion of the years saw negative or stagnant growth. In contrast, 75% shows the higher range of growth rates. For example, California, with 5.15%, shows that in three-quarters of the years, the growth rate was above this threshold. Lastly, the minimum and maximum values represent the extremes of income growth for each state.

The heterogeneity reflects a significant economic disparity between regions. In the first group (states such as Iowa, South Dakota, Washington, etc.), we can see high growth rates with minimal fluctuations, suggesting that these states may have seen steady economic expansion. On the other hand, states like Kentucky, Georgia, West Virginia show low or even negative growth at times, indicating economic struggles, stagnation or slow recovery after recessions. In reality, states like California, Texas, and Washington are benefiting from booming economies and are driven by high-wage sectors like technology and finance. Meanwhile, Mississippi, Kentucky and West Virginia are seeing stagnation. It seems that there are some industries that have faced decline. Moreover, states with larger negative growth in the years immediately following the 2008 financial crisis (e.g., Nevada, Arizona, Florida) may have been impacted by the downturn. However, by 2015, many of these states showed recovery. This could indicate that their local economies are beginning to rebound after the crisis. Meanwhile, states like Kentucky, Mississippi show lower income growth rates. This may show slower recovery or other economic issues. In summary, the heterogeneity in income growth rates highlights the disparities in economic performance.

### **Assumptions for data challenges**

1. Missing values

Rows with State\_abbr = 'PR' (which stands for Puerto Rico) in the mortgage dataset don't have corresponding values in the median income dataset. We need to handle missing data for Puerto Rico in some ways, such as filling with assumed median income for Puerto Rico (e.g., using the average of nearby years or an industry standard) or removing rows.

There are 350 rows with missing values in the state\_orig\_time column (0.5%), clean up by removing these rows.

For numerical columns, we can impute missing values using the mean, median, or interpolation methods. If missing values are for columns like payoff\_time or status\_time, they might indicate that no action was taken during that period, so we can fill them with 0.

2. State Matching

In merged data, the state\_orig\_time column corresponds to state abbreviations, while State\_abbr is the same in the merged data. If any state abbreviation mismatch occurs, we would need to perform a clean-up to ensure they match correctly across datasets.

3. Data Consistency

Ensure that Median Income and Year match correctly for each row. If there are inconsistencies, we might have to filter out rows where the data doesn't correspond to the expected mapping.

4. Year-to-Time Mapping

As the time ranges from 1 to 60, which corresponds to years between 2001 and 2015, there could be edge cases where the mapping might not align precisely due to gaps in the data or incorrect time values. We can handle this by checking if there is any time outside the expected range and correcting them as needed.

### **2. PD Modelling**

#### A. Estimate a basic credit risk model for mortgage default probabilities

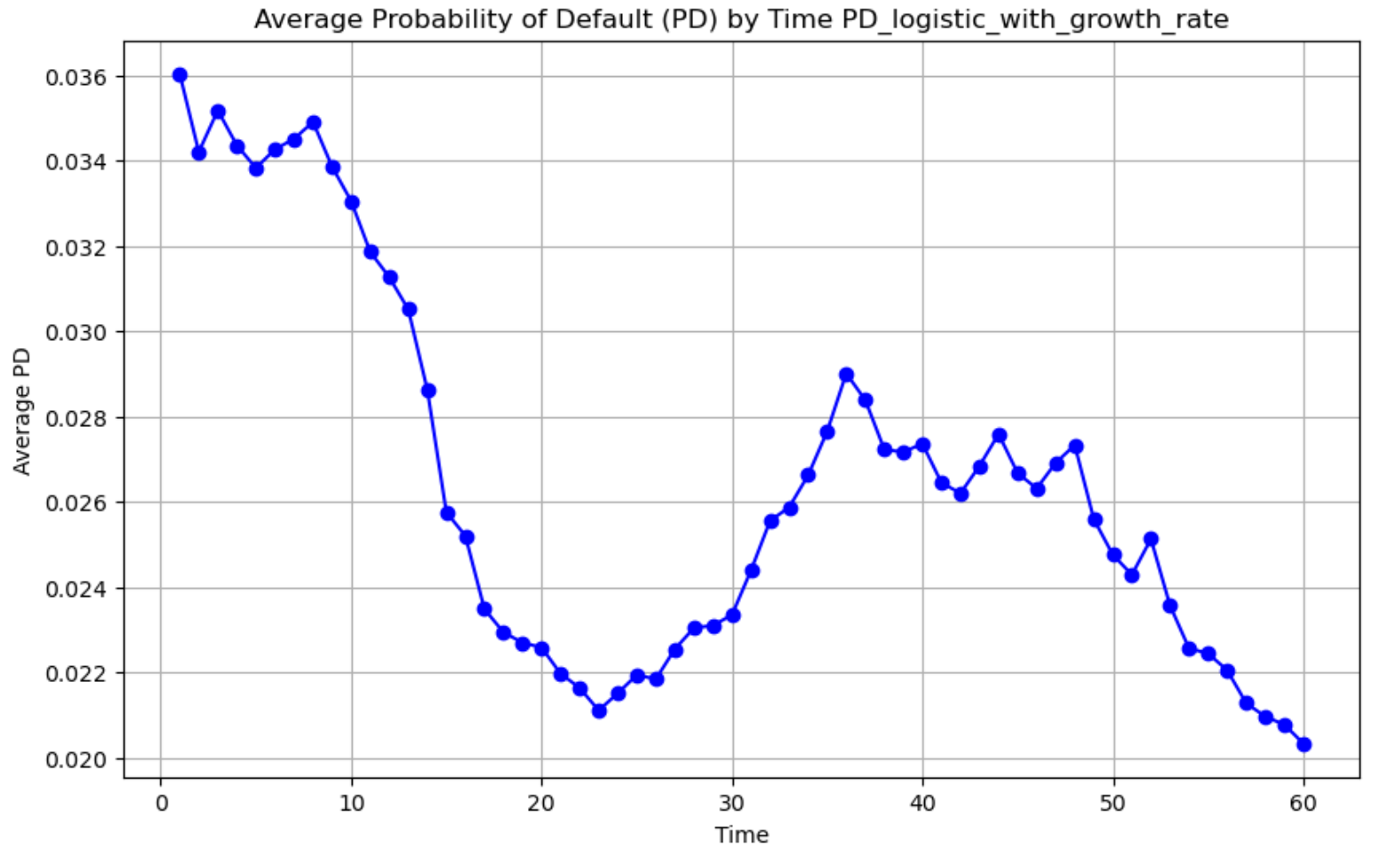
### 

### **Interpret the Plot**

The plot shows that the average probability of default fluctuates over time, peaking around time period 20 and then decreasing towards the end of the time period. This might suggest that borrowers have a higher likelihood of default at certain points in the mortgage life cycle, possibly linked to external factors like market conditions or borrower behavior.

Early spikes in PD indicate the beginning of economic distress, where borrowers are more likely to default due to macroeconomic challenges (e.g., economic recessions, high inflation, interest rate increases). Mid-periods show stabilization, suggesting recovery in income levels, financial conditions and possibly government interventions. The later periods demonstrate continued recovery with lower PD values indicating improved borrower repayment ability and economic stability. The sharp drop at the end shows that, after the economic distress, the situation improved, leading to fewer defaults.

#### B. Estimate the PD model again with level income growth



### **Interpret output**

The plot showing the Average Probability of Default (PD) by Time with State Income Growth provides insights into the relationship between the time period and the default probability of mortgage loans across different states, while also factoring in the income growth within those states.

Initial Periods (Time 0-10):

Firstly, we see a spike in the average PD. This could signify a period of economic uncertainty or a time when the loans were more likely to default. During this time, many states may have been experiencing stagnation or slower growth, causing a higher likelihood of defaults. The income growth at the state level could be negatively impacted during these periods, making it harder for borrowers to meet repayment schedules.

Middle Periods (Time 10-30):

From Time 10 onward, we observe a sharp decline in PD, especially between Time 15 to 20 (2004:Q3 to 2005:Q4). This could represent a recovery period where the economic environment improved, potentially driven by favorable income growth across several states.

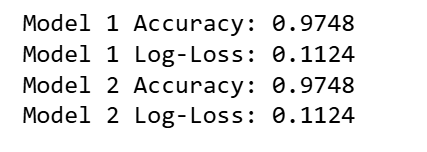
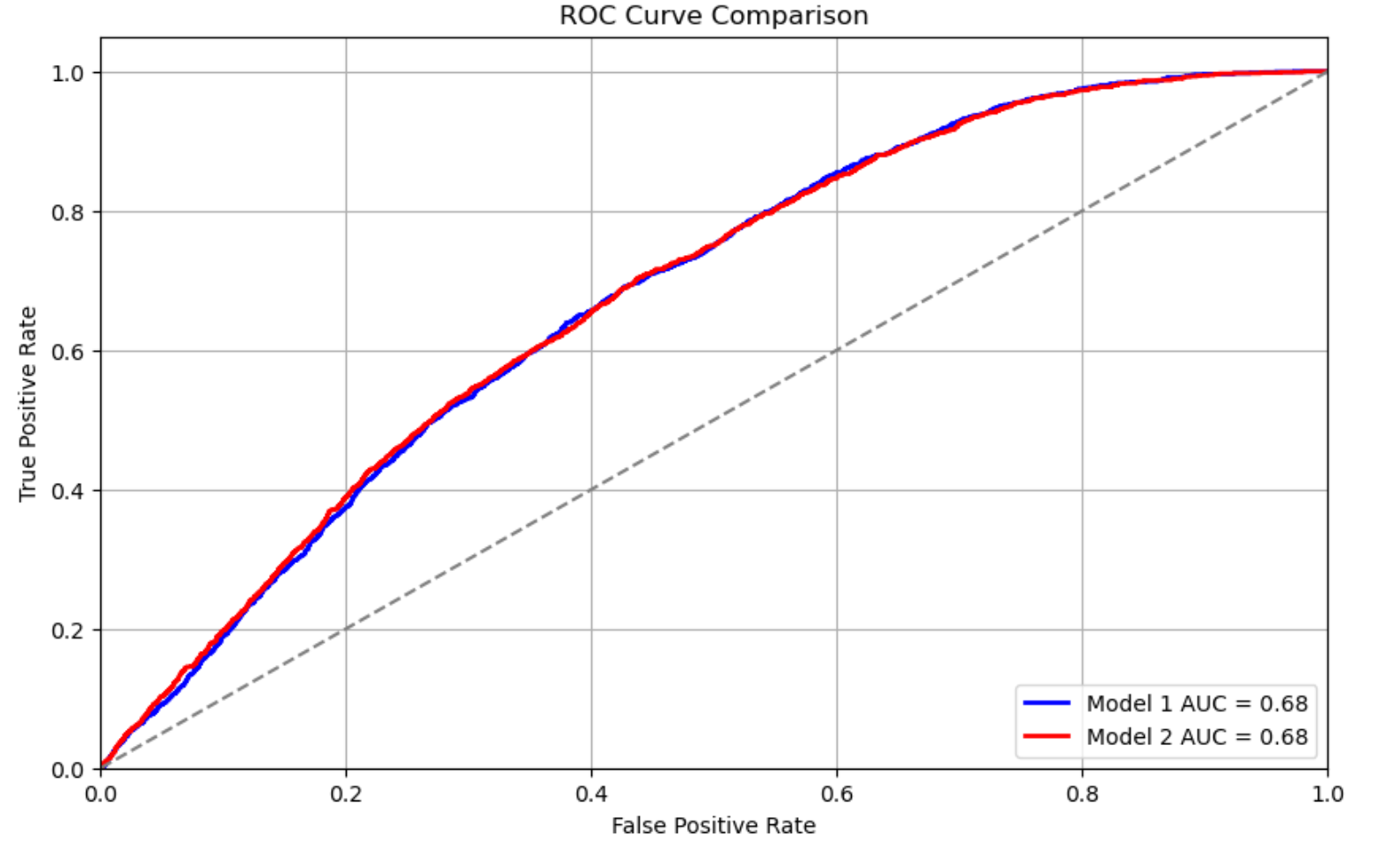
Later Periods (Time 30-60):

Flatter PD Curve shows that after Time 30, the PD stabilizes at a lower level, with a more gradual decline as time progresses. This might indicate that while some states have recovered from the economic downturn, the remaining loan defaults are mainly driven by other factors, such as changing interest rates or credit conditions.

Besides, there is a slight increase in PD after Time 50 (2012:Q2). This could suggest some new economic challenges, possibly related to external shocks such as recessions, tightening of credit, or regional economic slowdowns.

Overall, the relationship between income growth and default probability is inverse. When state-level income grows, the probability of default tends to decline. Conversely, when income stagnates or declines, the PD tends to increase as borrowers may struggle to meet payment schedules. The plot also suggests that macroeconomic factors, such as income growth, recession periods, etc. can affect mortgage defaults.

#### C. Compare the accuracy



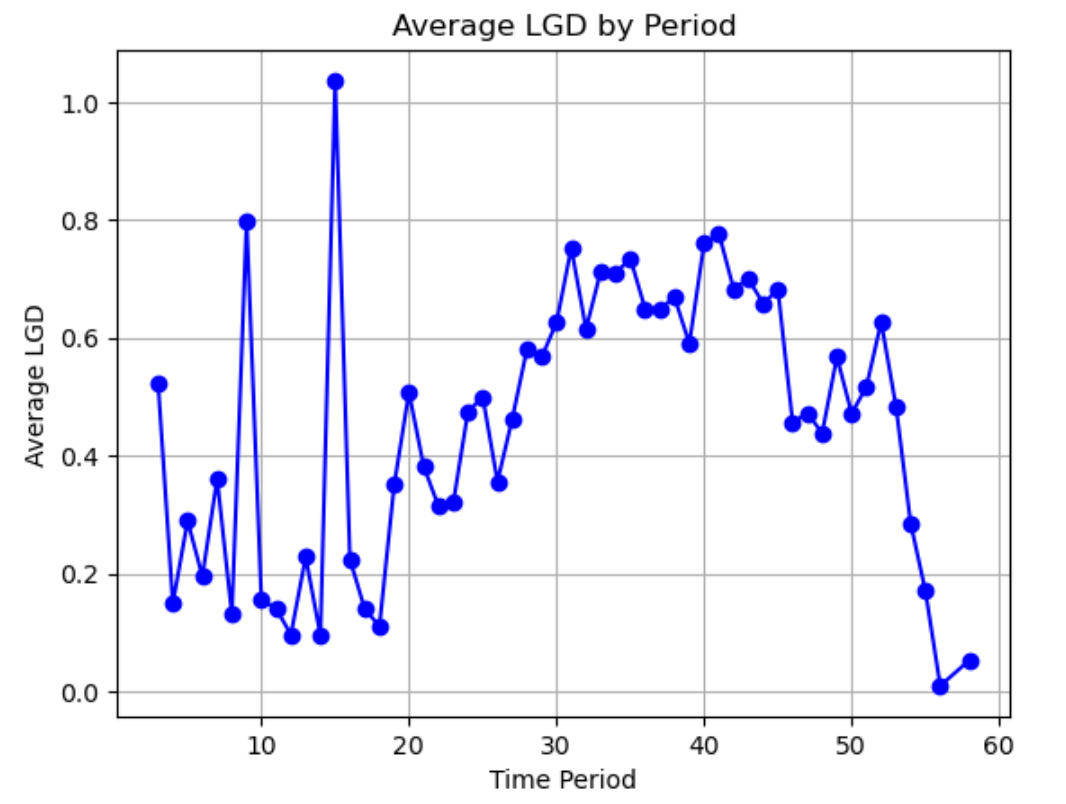
Model 1 and Model 2 perform similarly. However, Model 2's slight increase in AUC could be indicative of better classification performance at certain thresholds. The growth rate variable is potentially important but does not add significant value to the predictive power of the model.

Both models show a similar AUC score (Model 1: 0.6775 and Model 2: 0.6792), which indicates a similar ability to differentiate. The slight increase in AUC in Model 2 suggests that adding the growth rate variable slightly improves the model's ability to correctly classify the outcome, although the improvement is small.

Both models achieve an identical accuracy of 0.9748, which suggests that the accuracy rate is quite high. Both models have the same Log-Loss value of 0.1124. Log-Loss measures the uncertainty of the model’s predictions, with lower values indicating better model performance.

### **3. LGD Modelling**

#### A. Predict LGD



### **LGD Model**

The plot of Average LGD by Period shows how the Loss Given Default (LGD) varies over time across different periods.

The early periods (Time 1-10, namely 2001:Q1 - 2003:Q2) of high LGD reflect economic distress, likely tied to external shocks or systemic risks in the housing or credit markets. This volatility could also be due to data imbalances in the early periods or a smaller number of loans during this time frame. High LGD values indicate that, during these periods, borrowers were likely unable to repay loans due to poor economic conditions such as a downturn in the housing market, higher interest rates\*\*, or unfavorable income conditions at the state level.

There's a stabilization in the average LGD, especially between Time 10 and Time 30 (2003:Q2 - 2008:Q2). It suggests an improved financial environment, lower default rates, and better borrower repayment capacity. This may also reflect a recovery phase of the economy.

Although between Time 30 and Time 50 (2008:Q2 - 20013:Q2), the LGD remains relatively stable at lower levels, the late-period increases (Time 50-60) indicate that challenges such as rising inflation, credit tightening, or another economic downturn could be starting to affect borrowers' ability to repay loans.

### **4. Generative AI**

#### A. Additional Area: Portfolio Optimization in Credit Risk Prediction Process

Generative AI has promise for assisting the credit risk prediction process through increased productivity, prednisone and individualized services throughout the credit life cycle (McKinsey, 2024). There are various ways that generative AI might support including Data Augmentation, Feature, Anomaly Detection, Conditional Generation, Model Interpretation, Scenario Analysis and Fraud Detection. One additional area where generative AI can also greatly improve the process of predicting credit risk is portfolio optimization. According to a study by Moolchandani (2024), by simulating a variety of asset behaviors and market conditions, generative AI can help optimize portfolios. That helps financial institutions and banks to make more informed investment decisions, gain a better understanding of risk-return profiles and better manage credit portfolios. Particularly, generative AI can propose the best and optimal asset allocations throughout analyzing current and historical data. This assists in risk reduction strategy recommendation, real-time portfolio monitoring, default risk prediction and credit pricing setting (McKinsey, 2024).

In contrast, other generative AI applications focus on more specialized functions supporting the credit risk prediction process. First, data augmentation creates synthetic samples to counteract data scarcity and class imbalance to improve model robustness. Second, feature engineering generates new variables revealing hidden patterns in order to improve prediction accuracy without directly affecting asset allocations (Moolchandani, 2024). Third, anomaly detection using generative models to find outliers or unusual behavior is mainly used for fraud protection and data quality assurance. Conditional generation, which easily simulates particular borrower outcomes under predetermined economic scenarios is mostly used for scenario testing rather than allocation. The next area generative AI supports in the credit risk prediction process is scenario analysis, in which AI is used to stress test portfolios under simulated macroeconomic circumstances that provide information to guide but not carry out portfolio adjustments (Ajay, 2024). Lastly, fraud detection prioritizes financial security over optimization by using synthetic case generation and pattern learning to identify questionable activities (Stout, 2025). To sum up, portfolio optimization uses real-time intelligence to minimize risk across credit portfolios and maximize returns. In contrast, the other areas are supportive to enhance data quality, model performance, interpretability, and systemic risk awareness instead of directly controlling portfolio structure.

#### B. Scenario Generation: Forward-Looking Assessment of Income Growth and Credit Risk (2015–2029)

Selected Feature: State-Level Income Growth Rate

Metric Used: Annual percentage change in median household income

Application: Incorporated as a predictor in the PD model (Part 2B) and found to be statistically significant in the LGD model (Part 3A)

1. Historical Trend Analysis (2015–2023)

Between 2015 and 2019, U.S. states experienced moderate income growth, generally ranging between 2% and 3%, in line with a broader post-recession economic expansion. This period corresponded with reduced default probabilities and more favorable credit conditions. However, the emergence of the COVID-19 pandemic in 2020 leads to income stagnation or decline in several regions. Despite nominal wage increases in 2022–2023, elevated inflation effectively neutralized real income gains for many households.

Empirical results from Part 2B support this relationship. During periods of reduced or negative income growth (e.g., Time 0–10), the average probability of default (PD) peaked at over 0.034. Conversely, during Time 15–30, coinciding with rising income, PD declined significantly to approximately 0.022. In Part 3A, a similar relationship is observed for the Loss Given Default (LGD), where periods of improving income align with lower and more stable LGD values. Regression analysis further substantiates this association, revealing a statistically significant negative coefficient for income growth on LGD (–0.0902, p < 0.01).

2. Forward-Looking Scenario (2024–2029)

The following macroeconomic scenario has been constructed to simulate plausible income trajectories over the next six years. The scenario assumes global disinflation, gradual recovery in labor markets, and sectoral wage growth driven by technological adoption. These assumptions are consistent with recent central bank outlooks and OECD macroeconomic projections.

|  |  |  |
| --- | --- | --- |
| **Year** | **Economic Assumption** | **Projected Income Growth (%)** |
| 2024 | Persistently high interest rates, slow wage adjustment | 1.8% |
| 2025 | Declining inflation and initial labor market recovery | 2.5% |
| 2026 | Productivity gains in technology-intensive industries | 3.2% |
| 2027 | Broader wage expansion across lagging geographic regions | 3.5% |
| 2028 | Economic stabilization and normalization | 3.3% |
| 2029 | Mild cyclical slowdown with renewed credit tightening | 2.8% |

3. Projected Implications for Credit Risk

* Probability of Default (PD): According to the model findings in Part 2B, higher income growth rates are associated with lower default probabilities. Under the projected scenario, PD is expected to decline steadily from ~0.026 in 2024 to below 0.022 by 2028, provided income growth exceeds 2.5%. This trajectory mirrors the PD behavior observed in historical Time Periods 15–30.
* Loss Given Default (LGD): The LGD model (Part 3A) demonstrated that income growth significantly reduces the severity of losses following default. With sustained wage improvements, LGD is expected to remain within a stable range of 0.65–0.70 through 2028. However, in 2029, a slight uptick in LGD may occur, potentially rising to ~0.72—if income growth moderates alongside credit tightening, consistent with LGD trends observed between Time 50–60.

4. Data Source and Methodology

* Historical Data (2015–2023): Extracted from the U.S. Census Bureau’s H08 time series, as implemented in Part 2B.
* Projection Method: The scenario was generated using a hybrid approach combining historical trend extrapolation and generative AI-assisted simulation. It incorporates macroeconomic assumptions relating to inflation, labor dynamics, and fiscal policy, aligned with generative scenario design principles.
* Ratio Used: Income growth rate applied uniformly across state-level data to maintain consistency with PD and LGD model specifications.

This scenario demonstrates how income growth projections, when integrated with historical trends and economic foresight, can provide valuable input for credit risk modelling. The application of generative AI in scenario construction enhances the granularity and realism of assessments. This, in turn, supports more informed decision-making in areas such as stress testing, credit pricing and risk management.

### **5. Stress testing**

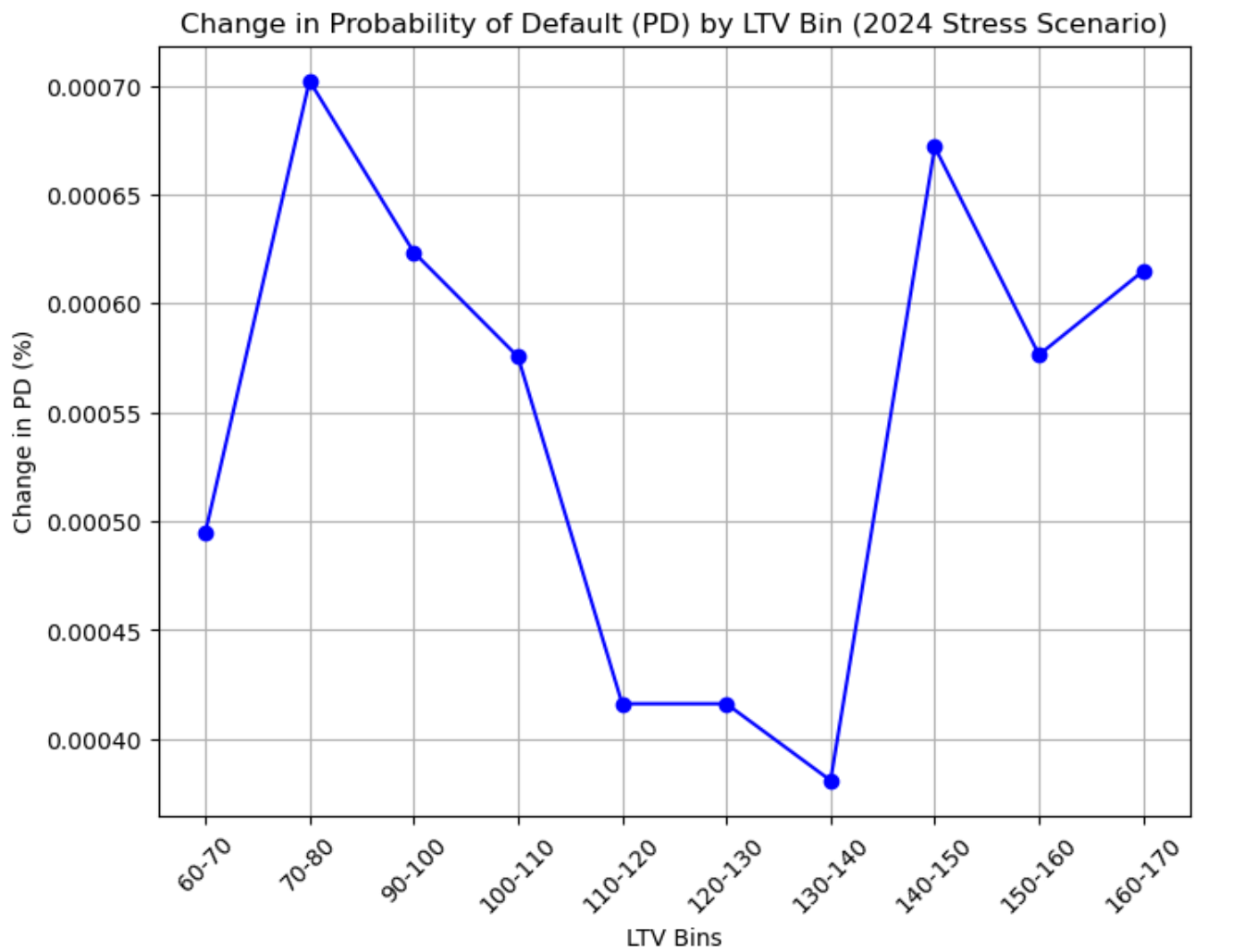
#### A. The change of PDs

The selected scenario is 2024 with 1.8% projected income growth, combined with 10 bins of original LTV ratio.

Baseline Period: This is the period where income growth was higher than the projected scenario (income growth of 2.5% or higher in earlier years like 2022–2023)

Stress Test Period: This will be the 2024 scenario, where projected income growth is 1.8%, reflecting economic challenges.

To conclude, the change in PD varies by LTV bins. The overall trend indicates greater sensitivity to economic conditions in certain LTV ranges, especially in the low and high ends.



#### B. Macroeconomic factors corresponding with double-trigger hypothesis in credit risk.

The first condition refers to a borrower’s inability to pay (often due to income loss). The macroeconomic factor corresponding with this hypothesis is rising unemployment. High unemployment leads to sudden income loss, the borrowers are unable to repay their loan. Borrowers facing job loss or reduced income are more likely to default if their home is also worth less than the mortgage balance (Pavan et al., 2020).

The second trigger is negative equity (when the mortgage exceeds the home’s value). This is related to housing market downturn. Falling home prices reduce the value of the collateral securing mortgages. When home values drop below the outstanding mortgage, borrowers experiencing financial distress are more likely to default. This not only reduces the ability to sell or refinance, but also makes it impossible for the borrower to continue paying if they found hopeless to recover their equity.

### **Reference**

1. Ajay, K. (2024). *Leveraging Gen AI in reimagining Credit Risk Management | Risk & Compliance Platform Europe*. Risk & Compliance Platform Europe. Retrieved from: <https://www.riskcompliance.biz/news/leveraging-gen-ai-in-reimagining-credit-risk-management/>
2. Barasa, I., Wanyonyi, S., & Kololi, M. (2025). *Application of Logistic Regression in Enhancing Digital Credit Risk Management in Commercial Banks*. Asian Journal of Probability and Statistics. Retrieved from: <https://doi.org/10.9734/ajpas/2025/v27i2710>.
3. Stout, D. W. (2025). *Generative AI for Financial Risk Prediction: Use cases. Maga*i. <https://magai.co/generative-ai-for-financial-risk-prediction-use-cases/>
4. McKinsey. (2024). *Embracing generative AI in credit risk*. McKinsey & Company. Retrieved from: <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/embracing-generative-ai-in-credit-risk>
5. Moolchandani, S. (2024). *The integration of generative AI in credit risk management*. www.academia.edu. Retrieved from: <https://www.academia.edu/124189188/The_Integration_of_Generative_AI_in_Credit_Risk_Management>
6. University of Technology Sydney (2025). *25751 Financial Institution Management. Lecture 1: Personal Lending.*
7. University of Technology Sydney (2025). 25752 Bank Lending and Analytics. Lecture 12: Credit Control
8. Pavan, M., & Barreda-Tarrazona, I. (2020). Should I default on my mortgage even if I can pay? Experimental evidence. Journal of Economic Dynamics and Control. <https://doi.org/10.1016/J.JEDC.2019.103733>.