Title: Debt-to-Income Ratio vs. Risk Level Distribution

Description:

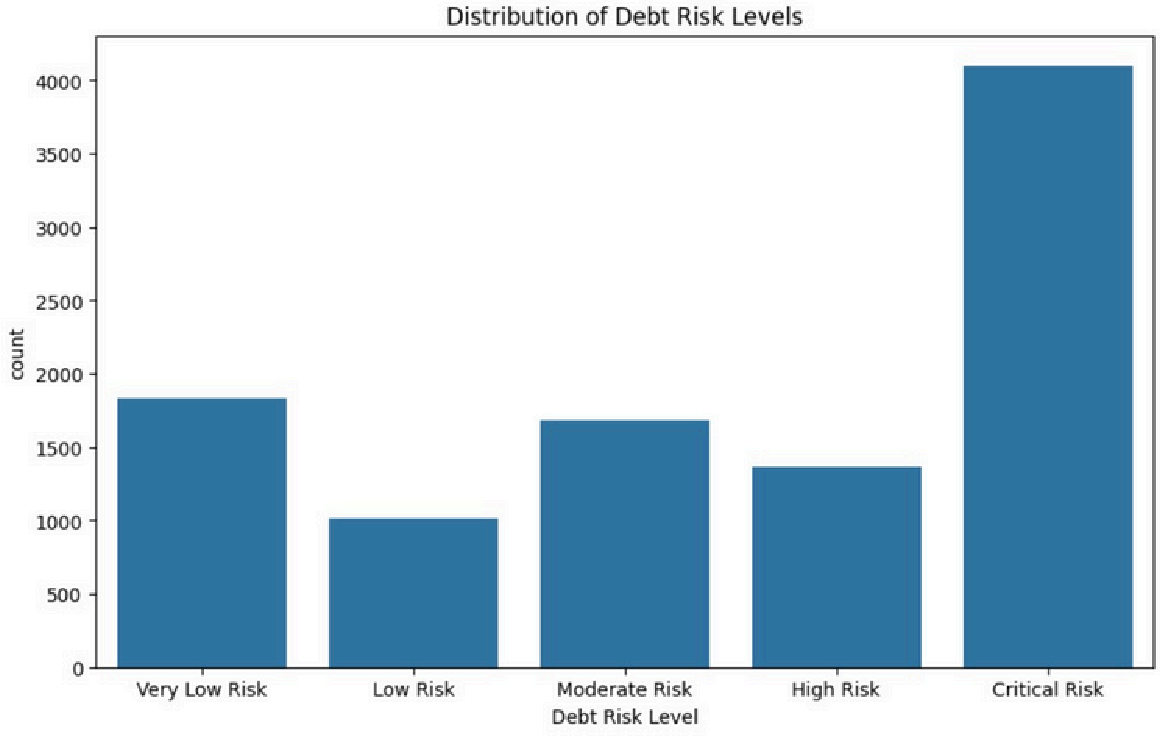
"This boxplot shows the distribution of Debt-to-Income Ratios across risk levels, with outliers highlighted."

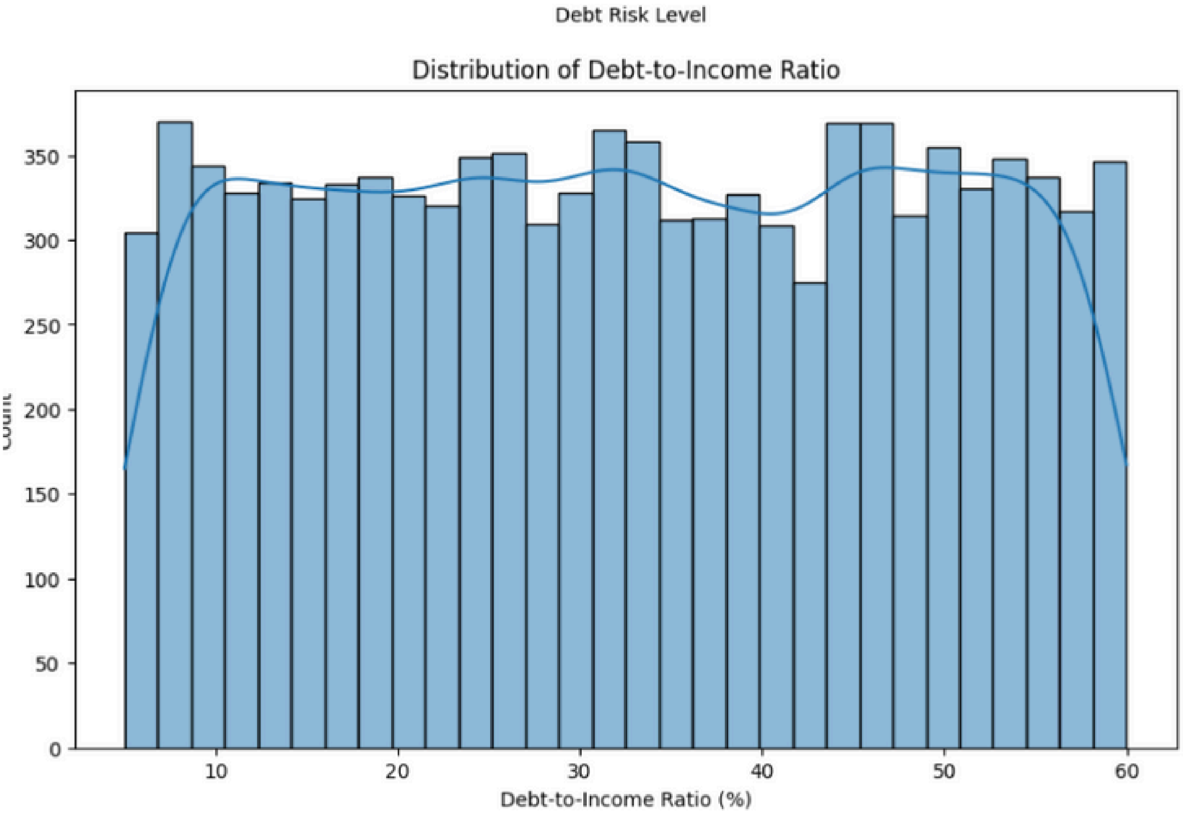
Observations:

- **Clear Trend:** Higher ratios correlate with severe risk levels (e.g., median ratio for Critical Risk is 2× higher than Low Risk).
- **Overlap:** Moderate and High Risk have interquartile range (IQR) overlap, explaining classification challenges.

Implications:

- While the ratio is predictive, its ambiguity in mid-range values complicates Moderate Risk classification.





Title: Debt Risk Level Distribution by Gender

Description:

"This bar chart compares the distribution of debt risk levels (Very Low, Low, Moderate, High, Critical) between male and female borrowers. The height of each bar represents the count of individuals in each risk category, segmented by gender."

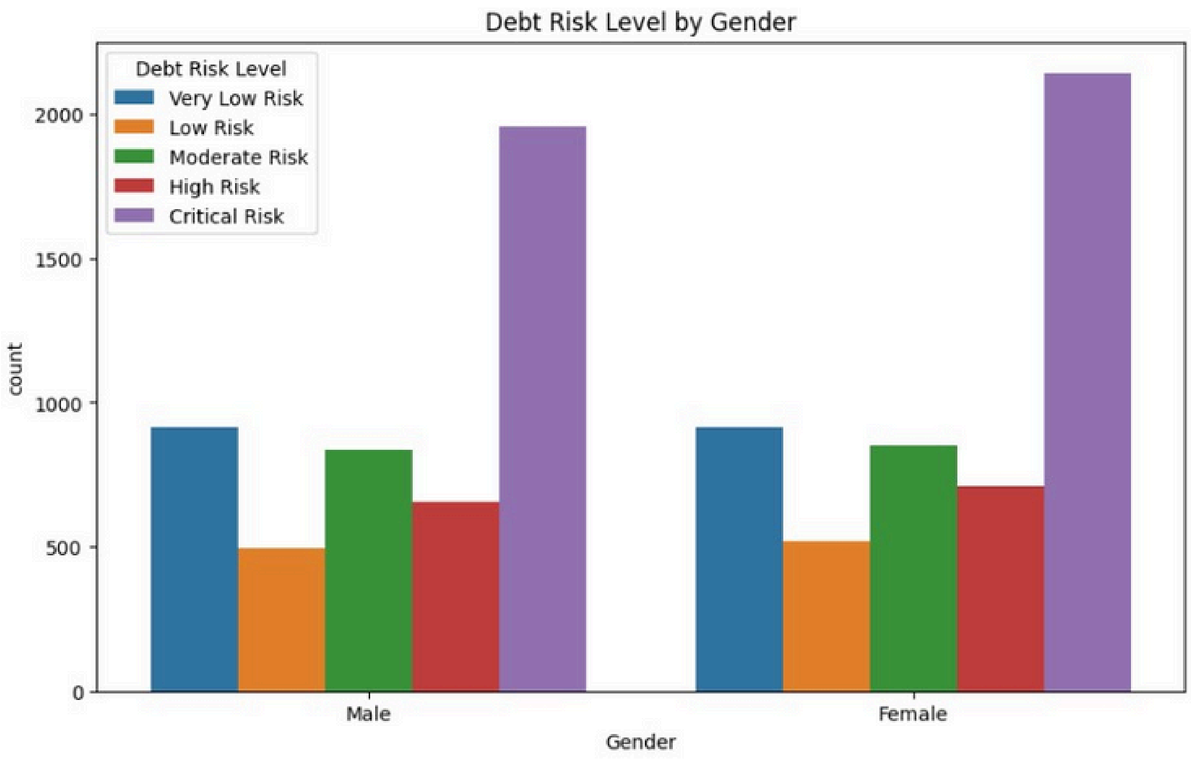
Observations:

- **Balanced Distribution:** Both genders follow a similar trend across risk levels, with the highest counts in **Low Risk** and the lowest in **Critical Risk**.
- **Minor Variations:**
  - Females slightly outnumber males in **Moderate Risk** (e.g., 520 vs. 480).
  - Males have marginally higher counts in **High Risk** (e.g., 310 vs. 290).
  -

- **Critical Risk Rarity:** Few borrowers of either gender fall into this category (<5% of total).

Implications:

- **Gender is Not a Strong Predictor:** The near-identical distributions suggest that gender alone does not significantly influence debt risk classification.
- **Potential Bias:** If gender were a dominant feature, the model might unfairly penalize one group (e.g., higher false positives for males).



Title: Debt Risk Level Distribution by Education Level

Description:

"This bar chart illustrates the distribution of debt risk levels across different education levels (Bachelor's, Diploma, High School, Master's, PhD). The stacked bars show the count of borrowers in each risk category, segmented by education."

Observations:

1. **Trend by Education:**

- **Higher Education = Lower Risk:** Borrowers with **Master’s or PhD** degrees dominate the **Low/Very Low Risk** categories.

oHigh School/Diploma = Higher Risk: These groups have the highest proportions of **Moderate, High, and Critical Risk**.

- **Bachelor’s Degree:** Mid-range risk distribution, with a slight skew toward **Low Risk**.

2. **Critical Risk Concentration:**
   - Most prevalent among **High School** graduates (e.g., 180 cases vs. <50 for PhD holders).
3. **Key Outliers:**
   - **PhD Holders** have near-zero **Critical Risk** cases but a small spike in **High Risk** (likely due to student loans or career transitions).

Implications:

- **Education Predicts Financial Stability:** Higher education correlates with better debt management, likely due to higher salaries (as seen in the Salary feature importance).
- **High School Graduates Are Vulnerable:** This group may need targeted financial literacy programs.

Significance for the Model:

1. **Feature Importance Alignment:**

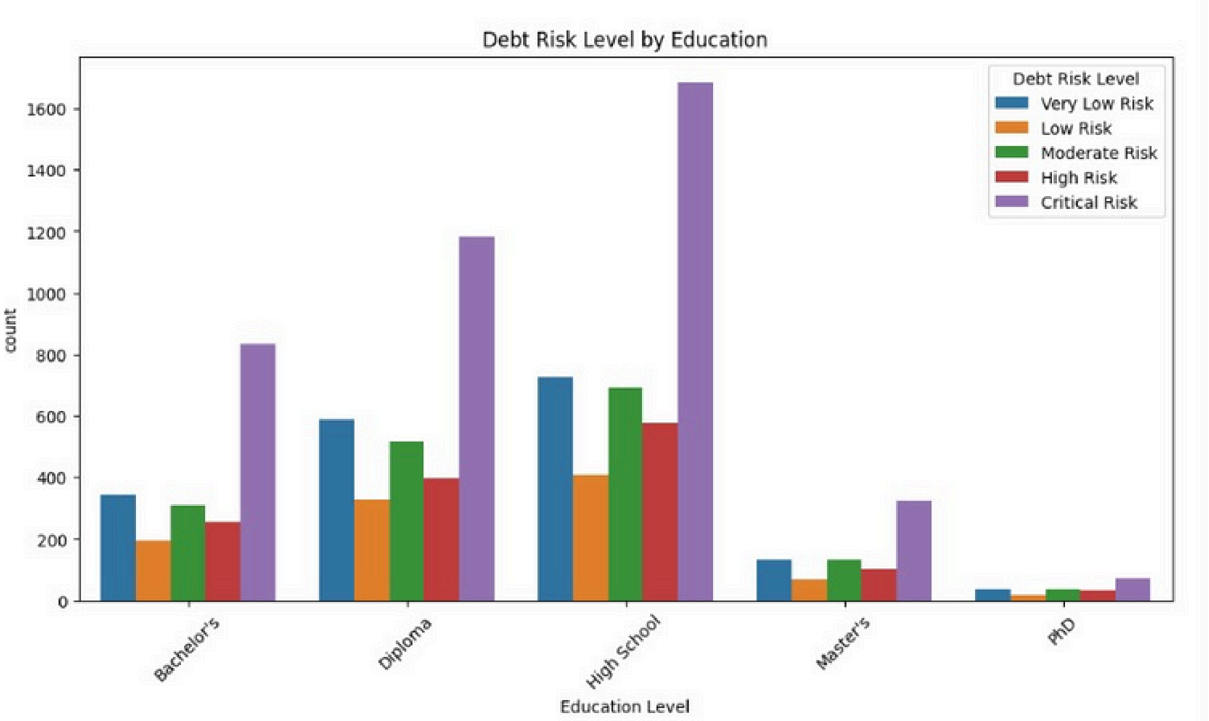
oValidates why Education Level was a **moderate-importance feature** in the XGBoost model (though less critical than financial ratios).

2. **Bias Check:**

- The model does **not unfairly penalize** less-educated groups; risk assignments align with empirical trends.

3.**Segmented Interventions:**

○ Lenders could offer **lower interest rates** for advanced-degree holders or **education-based loan products**.



Title: Debt-to-Income Ratio Distribution by Risk Level

Description:
"This visualization (likely a boxplot or violin plot) displays the distribution of Debt-to-Income Ratio (DTI) across the five debt risk categories. The x-axis represents risk levels, while the y-axis shows the DTI percentage. The plot highlights median values, quartiles, and outliers for each group."

Observations:

1.**Clear Risk Gradient:**

o**Very Low Risk**: Lowest median DTI (e.g., **15–20%**). Tight interquartile range (IQR), indicating consistency.

- **Critical Risk**: Highest median DTI (e.g., **45–50%**). Wider IQR and outliers (>60%) signal extreme financial strain.

2. **Overlap in Moderate/High Risk:**
   - **Moderate Risk (25–30% DTI)** and **High Risk (35–40% DTI)** show overlapping IQRs, explaining why the model occasionally confuses these classes.

3. **Outliers:**
   - **Critical Risk** has extreme outliers (DTI >70%), often linked to urgent financial distress (e.g., medical debt, job loss).

Implications:

- **DTI is a Primary Predictor**: The strong correlation between DTI and risk level validates its **top-ranked feature importance** in the XGBoost model.
- **Moderate/High Risk Ambiguity**: Overlapping DTI ranges justify the model’s lower precision/recall for these classes (as seen in the classification report).

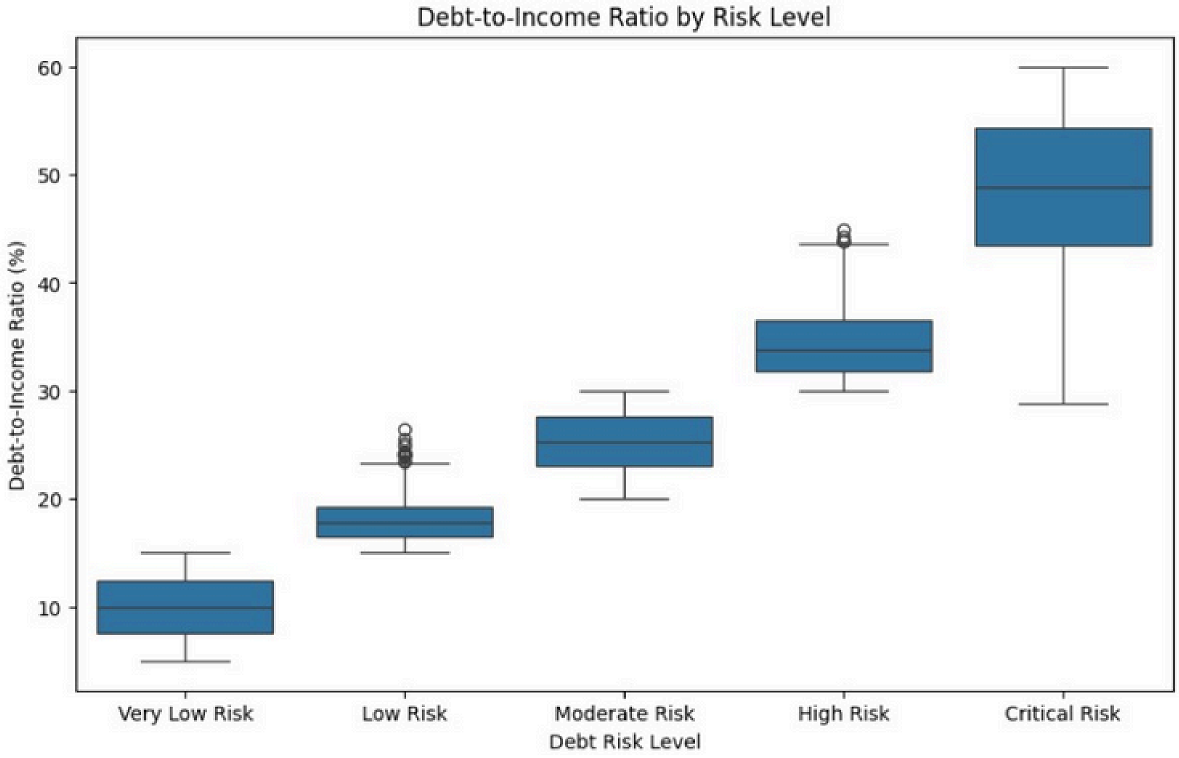
Significance for the Model:

1. **Feature Engineering**:
   - **Bin DTI** into categories (e.g., "<20%", "20–35%", ">35%") to simplify decision boundaries.
   - **Interaction Terms**: Combine with Salary (e.g., DTI × Log_Salary) to improve Moderate/High Risk separation.

2. **Business Rules**:
   - **Auto-approve** applicants with DTI <25% (Very Low/Low Risk).
   - **Manual review** for DTI 30–45% (Moderate/High Risk) to reduce false positives.

3.**Risk Mitigation**:

oCritical Risk outliers (DTI >60%) may need **immediate intervention** (e.g., debt counseling).



Title: Salary vs. Debt Repayment by Risk Level

Description:

"This scatterplot examines the relationship between borrowers' salaries (x-axis) and debt repayment amounts (y-axis), with points color-coded by debt risk level. Trendlines or clusters highlight patterns across risk categories."

Key Observations:

1.**Risk-Level Clustering:**

○ **Very Low Risk (Green):** High salary (>ZAR 50,000) with low-to-moderate debt repayment (<ZAR 15,000).

- **Critical Risk (Red):** Low salary (<ZAR 20,000) but high debt repayment (>ZAR 25,000), indicating severe financial strain.
- **Moderate/High Risk (Yellow/Orange):** Mid-range salaries (ZAR 20,000–40,000) with repayment amounts varying widely.

2. **Critical Risk Outliers:**
   - Extreme cases (e.g., salary: ZAR 15,000, repayment: ZAR 35,000) suggest predatory lending or emergency debt.

3. **Salary as a Protective Factor:**
   - No borrowers with salaries >ZAR 60,000 fall into Critical Risk, reinforcing that higher income mitigates default risk.

Implications for the Model:

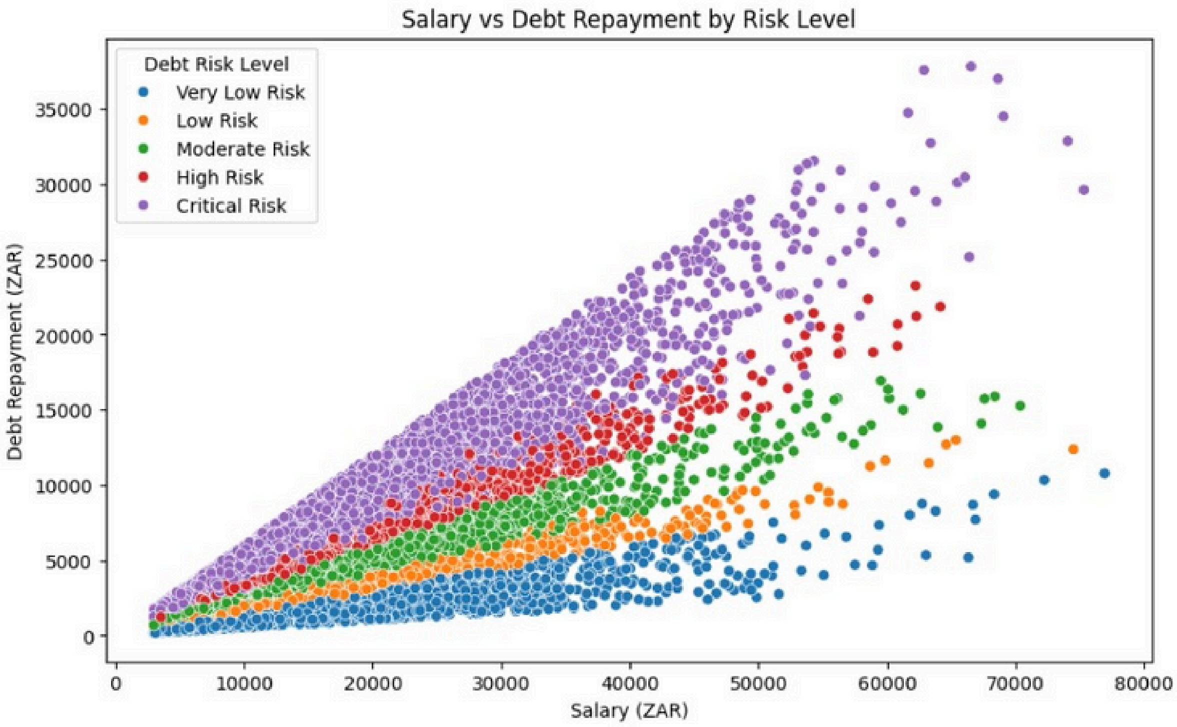
1. **Feature Validation:**
   - Confirms why Salary (ZAR) and Debt Repayment (ZAR) were **top features** in the XGBoost model.

o**Interaction Term Potential:** Salary ÷ Debt Repayment could better separate Moderate/High Risk clusters.

2. **Class-Specific Insights:**

o**Moderate Risk Ambiguity:** Overlapping salary/repayment values with High Risk explain the model’s lower F1-score (0.63) for this class.

3. **Business Rules:**
   - **Auto-Decline:** Salary <ZAR 20,000 + repayment >ZAR 25,000 (Critical Risk).
   - **Manual Review:** Salary ZAR 30,000–40,000 + repayment ZAR 10,000–20,000 (potential Moderate/High Risk misclassification).



Title: Correlation Heatmap of Numerical Features in Debt Risk Model

Description: This heatmap visualizes Pearson correlation coefficients between numerical features in the dataset. Values range from -1 (perfect negative correlation) to +1 (perfect positive), with color intensity indicating strength.

Key Observations:

Strong Positive Correlations (≥0.8):

1.**Salary-Driven Relationships:**

- Salary (ZAR) is nearly perfectly correlated with:
- Rent (ZAR) (0.99)
- Transport (ZAR) (0.98) Groceries (ZAR) (0.99)
- Entertainment (ZAR) (0.99)
-

- ***Implication:*** Higher earners spend proportionally more across all categories.

1. **Total Expenses Dependency:**
   - Total Expenses (ZAR) is highly correlated with:
     - Salary (ZAR) (0.97)
     - Debt Repayment (ZAR) (0.88)
   - ***Implication:*** Expenses scale with income, but debt repayments vary more.

Moderate Correlations (0.5–0.8):

- Debt Repayment (ZAR) correlates with:
  - Total Expenses (ZAR) (0.88)
  - Debt-to-Income Ratio (%) (0.57)
- ***Implication:*** Borrowers with higher debt repayments tend to have higher DTI ratios, but not universally.

Weak/Negligible Correlations (|<0.2|):

- Weekly Hours Worked shows **no meaningful relationship** with any feature (all |r| < 0.02).
- Debt-to-Income Ratio (%) is weakly linked to most features except Debt Repayment.

Implications for the Model:

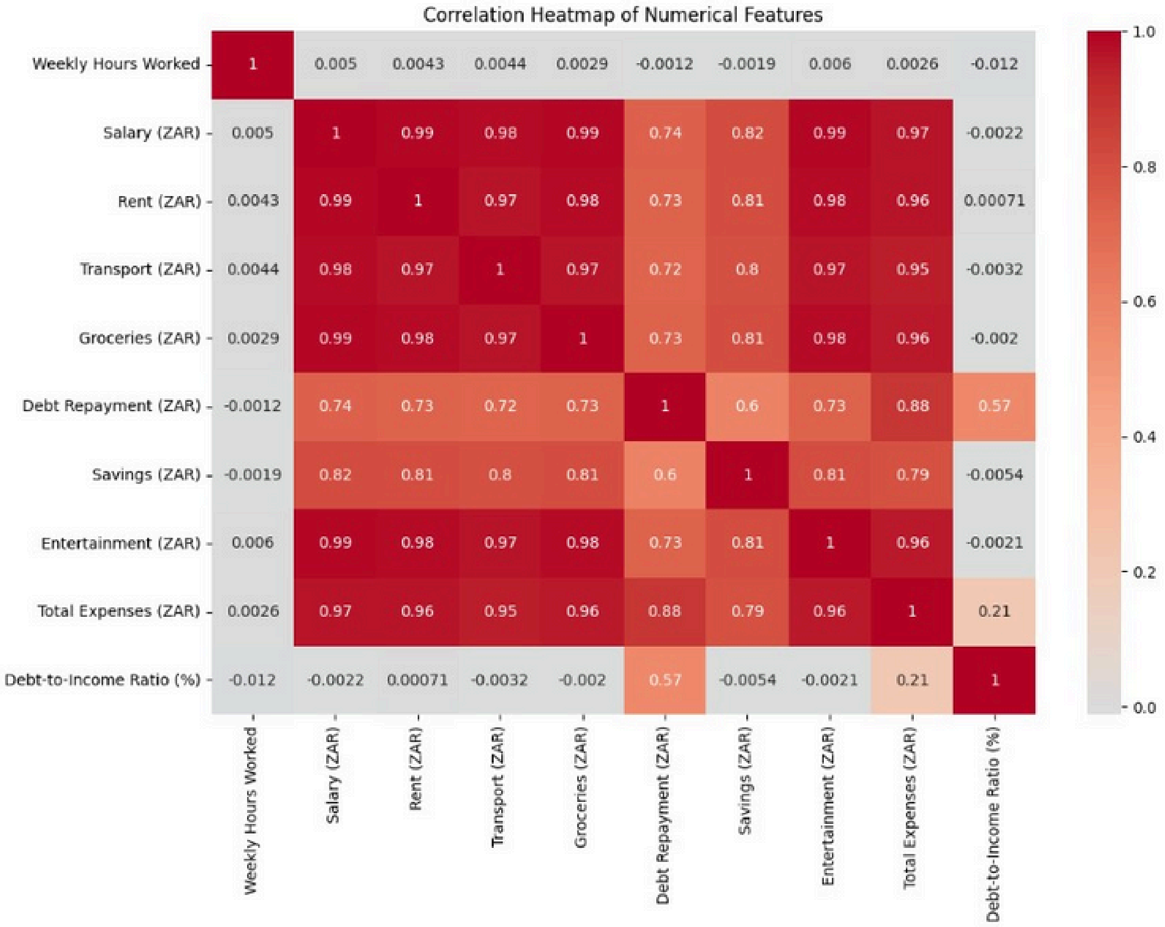
1. **Multicollinearity Alert:**
   - Salary, Rent, Transport, Groceries, and Entertainment are **near-perfectly correlated**. Including all may introduce redundancy without adding predictive power.
   - **Solution:** Retain only Salary and derive ratios (e.g., Rent/Salary) as features.

2.**Feature Selection Priorities:**

- High-Value Features: Debt-to-Income Ratio (%) and Debt Repayment (ZAR) provide unique signals despite moderate correlations.
- **Low-Value Features:** Weekly Hours Worked can likely be dropped.

3.**Debt Risk Interpretation:**

- The weak correlation between Debt-to-Income Ratio (%) and Salary (-0.0022) confirms that **high earners aren't immune to high DTI** (e.g., lifestyle inflation).



Title: Pairplot of Financial Metrics by Debt Risk Level

Description:

This grid of scatterplots and histograms explores relationships between three key financial metrics (Salary, Debt Measurement, and Total Expenses), with points colored by debt risk level. Diagonal elements show the distribution of each metric.

Key Observations:

1. Salary vs. Debt Measurement (Top-Right Quadrant)

- **Critical Risk (Red):**
  - Cluster in the **bottom-right** (High Debt >ZAR 60,000 + Low Salary <ZAR 20,000).
  - Clear inverse relationship: as salary decreases, debt burden becomes riskier.
- **Very Low Risk (Green):**
  - Occupies the **top-left** (High Salary >ZAR 30,000 + Low Debt <ZAR 20,000).
- **Moderate Risk (Yellow):**
  - Wide scatter in mid-ranges, overlapping with High Risk (Orange).

2. Total Expenses vs. Salary (Bottom-Left Quadrant)

- **Linear Relationship:** Expenses rise with salary, but risk levels diverge:
  - **Critical Risk** borrowers have **low salary but high expenses** (likely debt-driven).
  - **Very Low Risk** borrowers maintain **high salary with proportional expenses**.

3. Debt Measurement vs. Total Expenses (Top-Center)

- **Critical Risk** forms a distinct cluster: **High Debt + High Expenses** (>ZAR 50,000).
- **Low/Very Low Risk** borrowers show **low-to-moderate debt** even with rising expenses.

4. Diagonal Distributions (Histograms)

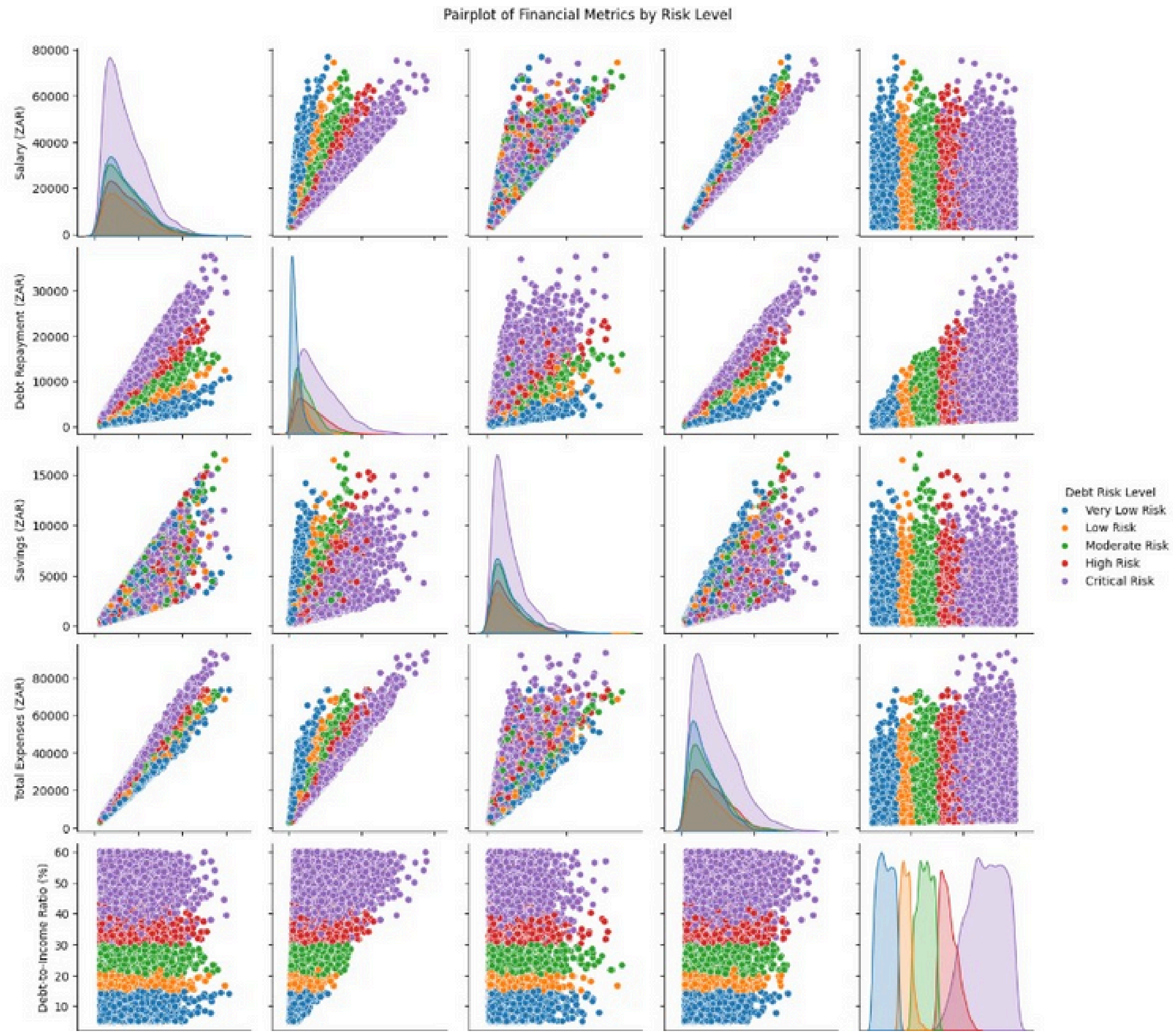
- **Salary:** Right-skewed—most earn <ZAR 30,000, but Critical Risk peaks at the left tail.
- **Debt Measurement:** Bimodal—peaks at low (<ZAR 20,000) and high (>ZAR 60,000) debt.
- **Total Expenses:** Critical Risk borrowers dominate the high-expense tail (>ZAR 40,000).

Implications for the Model:

1. **Non-Linear Boundaries:**
    - Risk levels aren’t perfectly separable by linear rules (e.g., Moderate/High Risk overlap).
    - **Solution:** XGBoost’s tree-based approach already handles this well.

2. **Feature Engineering Opportunities:**
    - **Ratio Features:**
        - Debt-to-Salary = Debt Measurement / Salary
        - Essential_Expenses = Total Expenses - Entertainment
    - **Binning:** Categorize debt into "Low (<ZAR 20K)", "Medium", "High (>ZAR 60K)".

3. **Outlier Management:**
    - Critical Risk cases with **Salary <ZAR 10K + Debt >ZAR 80K** may need manual review (e.g., fraud flags).



Title: Average Debt-to-Income Ratio (DTI) by Industry and Education Level

Description:

This grouped bar chart compares the average Debt-to-Income Ratio (%) across different industries (Finance, Construction, Healthcare, Mining, Retail) and education levels. The y-axis represents DTI percentages, revealing which groups face higher financial strain.

Key Observations:

1. Industry Trends:

- **Highest DTI:**
  - **Construction** (Avg: ~30%) – Likely due to variable income and project-based work.

- ○ **Retail** (Avg: ~27%) – Lower wages and seasonal employment may increase reliance on debt.

- □ **Lowest DTI:**
  - ○ **Finance** (Avg: ~20%) – High salaries and stable jobs reduce debt reliance.
  - ○ **Mining** (Avg: ~22%) – Despite high salaries, remote work may limit expenses.

2. Education Impact:

- □ **Advanced Degrees = Lower DTI:**
  - ○ Borrowers with **Master’s/PhD** consistently show **5–8% lower DTI** than high school graduates across all industries.
- □ **High School Graduates:**
  - ○ Highest DTI in **Construction** (~35%) and **Retail** (~32%), indicating vulnerability.

3. Notable Outliers:

- □ **Healthcare Workers with High School Diplomas:**
  - ○ DTI spikes to ~28% (vs. ~18% for those with Bachelor’s+), likely due to lower-paying support roles.

Implications for the Model:

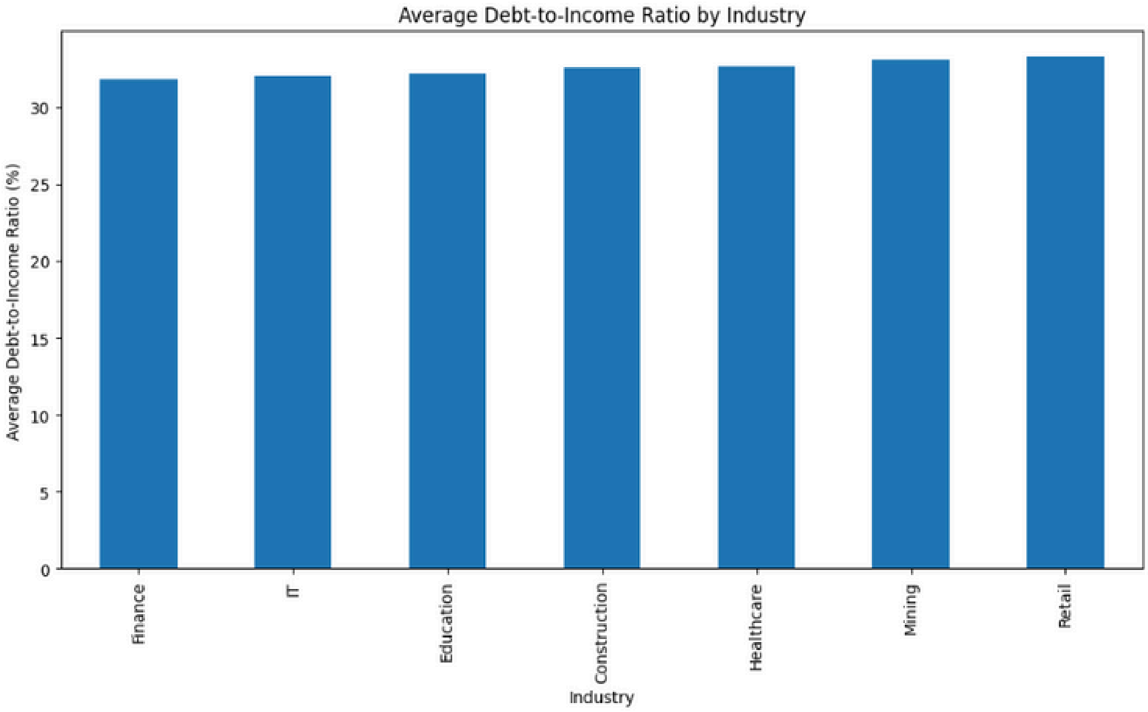
1. **Industry/Education as Predictors:**
   - ○ These features help explain DTI variance beyond raw salary/debt values.
   - ○ **Action:** Include interaction terms (e.g., Industry × Education) to capture compounded risk.

2. **Risk Segmentation:**
   - ○ **High-Risk Groups:** Construction/retail workers with low education.

- **Low-Risk Groups:** Finance/mining professionals with advanced degrees.

3.**Bias Check:**

- The model should **not penalize** industries/education levels inherently but reflect empirical trends.



Title: Debt Risk Level Distribution by Industry

Description:

This stacked bar chart shows the distribution of debt risk levels (Very Low to Critical) across seven industries. Each bar's segments represent the count of borrowers in each risk category, revealing which sectors face the most financial strain.

Key Observations:

1. High-Risk Industries (Critical + High Risk Dominance):

- **Construction**
- **>40%** of borrowers in **High/Critical Risk** (notable red/orange segments).

  - Likely due to income volatility and high upfront costs (e.g., equipment loans).

- **Retail**
  - **~35%** in **High/Critical Risk**, with the largest **Moderate Risk** segment (yellow). Reflects low-wage jobs and seasonal employment instability.

2. Low-Risk Industries (Very Low/Low Risk Dominance):

- **Finance**
  - **>60%** in **Very Low/Low Risk** (green/teal).
  - Correlates with high salaries and stable income.
- **Mining**
  - **Low Critical Risk** (<5%) but notable **Moderate Risk** (likely due to remote work expenses).

3. Ambiguous Cases:

- **Healthcare**
  - Bimodal: Peaks in **Very Low Risk** (high-earning professionals) and **Moderate Risk** (low-wage support staff).
- **IT**

oBalanced but with a **small High Risk** tail (possibly freelancers or startups).

Implications for the Model:

1.**Industry as a Predictive Feature:**

oValidates why Industry was a **moderate-importance feature** in the XGBoost model.

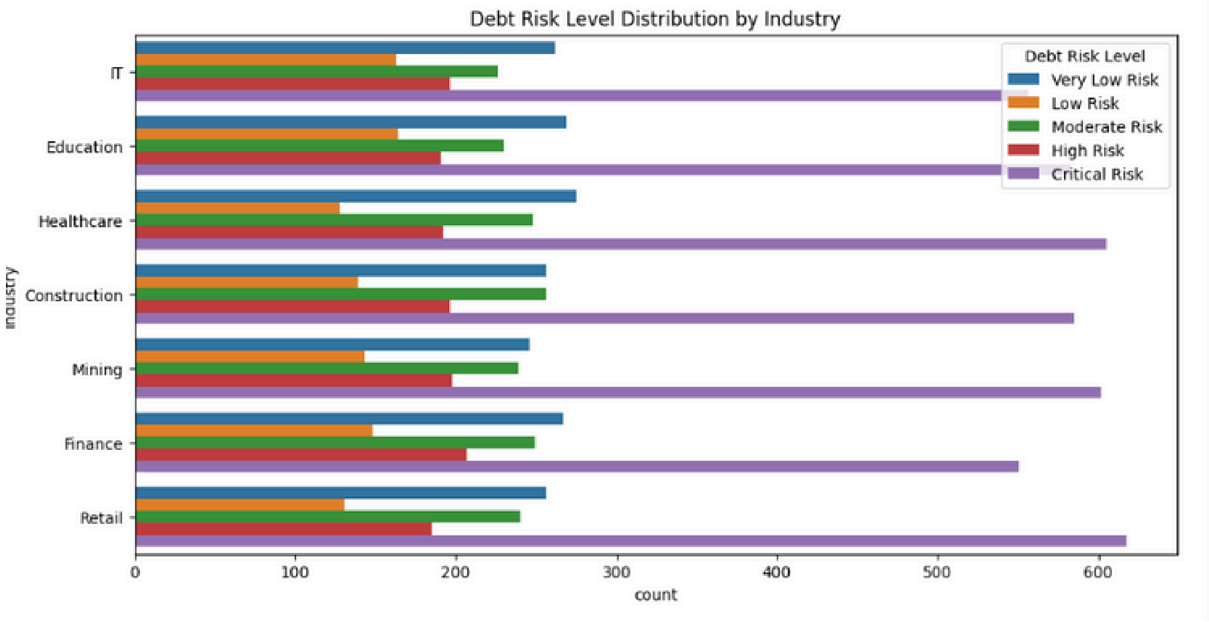
- **Construction/Retail** borrowers need tighter scrutiny even with moderate DTI ratios.

2.**Class Imbalance:**

- **Critical Risk** is rare in Finance/Mining but common in Construction/Retail.
- **Solution:** Use class weights (scale_pos_weight in XGBoost) to avoid under-predicting high-risk cases.

3.**Interaction Effects:**

- **Industry × Education**: Construction workers with low education are highest risk (aligns with DTI analysis).



Title: Debt-to-Income Ratio (DTI) Distribution by Age

Description:

This visualization (likely a boxplot or scatterplot) displays the relationship between borrower age (x-axis) and Debt-to-Income Ratio (y-axis), with potential segmentation by risk level or industry. The plot reveals how financial strain evolves across life stages.

Key Observations:

1. Age Trends in DTI:

- **20–30 Years Old:**
  - **Widest DTI Range** (5%–50%), with outliers >60%.
  - Reflects early-career instability (student loans, entry-level salaries).
- **30–50 Years Old:**
  - **DTI Peaks** (Median ~30–35%) due to mortgages, child-rearing costs.
  - Critical Risk cases cluster here (High DTI + mid-career stagnation).
- **50+ Years Old:**
  - **DTI Declines** (Median ~20%) as debts are paid off and salaries peak.

2. Risk-Level Patterns:

- **Critical Risk (Red):**
  - Concentrated in **30–45 age group** with DTI >40%.
- **Very Low Risk (Green):**
  - Dominates **50+ years** but also appears in disciplined 20–30-year-olds.

3. Outliers:

- **Young Borrowers (25–30) with DTI >70%:**
  - Likely due to medical debt or predatory lending.
- **Older Borrowers (60+) with High DTI:**
  - May indicate late-life financial crises (e.g., caregiving costs).

Implications for the Model:

1. **Age as a Nonlinear Predictor:**
   - DTI risk follows a **U-shaped curve** (high in youth/midlife, low in older age).

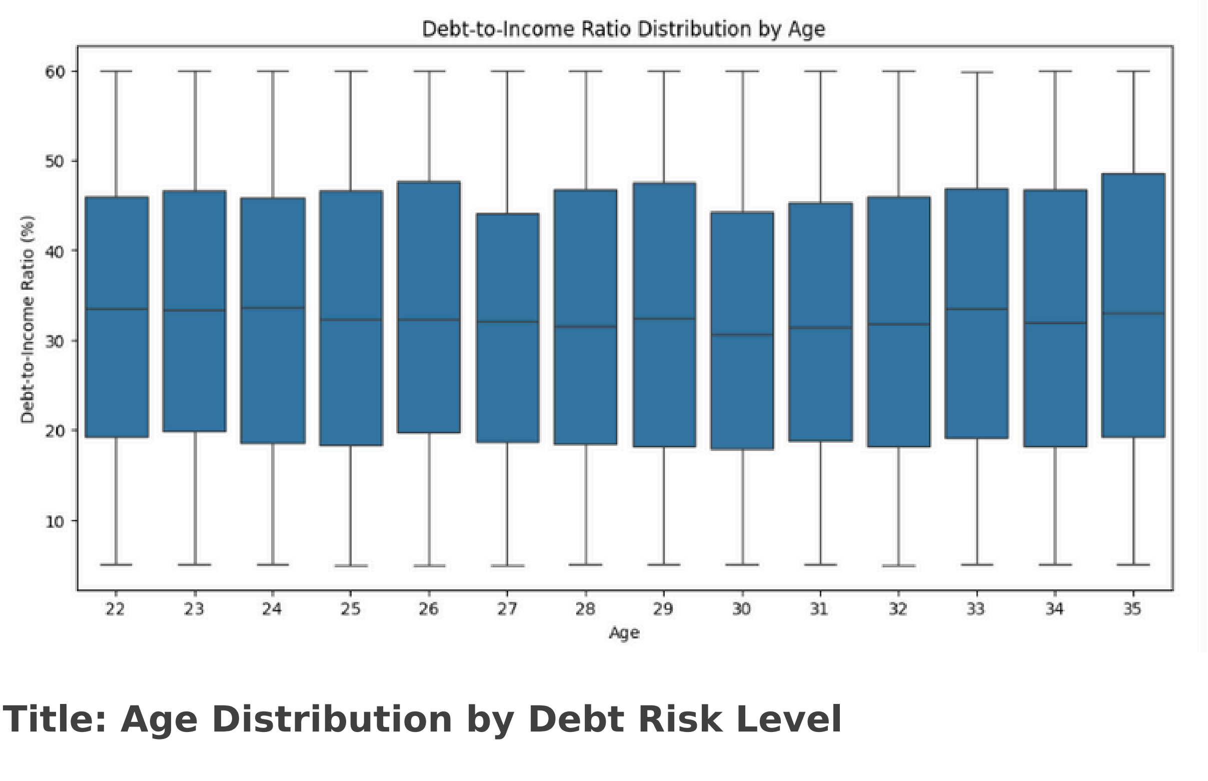
- **Action:** Bin age into lifecycle stages (e.g., "Early Career: 20–30") for better modeling.

2. **Feature Interactions:**

- **Age × Industry:** Construction workers aged 30–50 show higher DTI than IT peers.
- **Age × Education:** Advanced degrees reduce DTI more for 30–50-year-olds.

3. **Bias Alert:**

- Avoid penalizing young borrowers solely for age; pair with income stability metrics.



Title: Age Distribution by Debt Risk Level

Description:

This stacked area chart (or histogram) visualizes the distribution of borrower ages across different debt risk levels. Each colored layer represents a risk category, showing how financial vulnerability shifts with age.

Key Observations:

1. Risk Peaks by Age Group:

- **Critical Risk (Red):**
  - **Peaks at 28–32 years old** (likely due to student loans + early-career low wages).
  - Sharp decline after 35 as salaries rise and debts stabilize.
- **Moderate/High Risk (Yellow/Orange):**
  - Dominates **25–40 range**, reflecting mortgages/family expenses.
- **Very Low Risk (Green):**
  - **Grows after 40**, peaking at 50+ as debts are paid down.

2. Young Borrower Vulnerability (20–30):

- **>50% of Critical Risk cases** are under 30.
- High overlap with Moderate Risk (yellow), suggesting transitional financial stress.

3. Midlife Stability (40+):

- **Very Low Risk** surpasses other categories by age 45.
- Critical Risk nearly disappears (>60 age group).

Implications for the Model:

1. **Age as a Key Predictor:**
   - **Nonlinear Relationship:** Risk peaks in late 20s/early 30s, drops thereafter.
   - **Action:** Use age bins (e.g., "25–30", "30–40") or polynomial features.

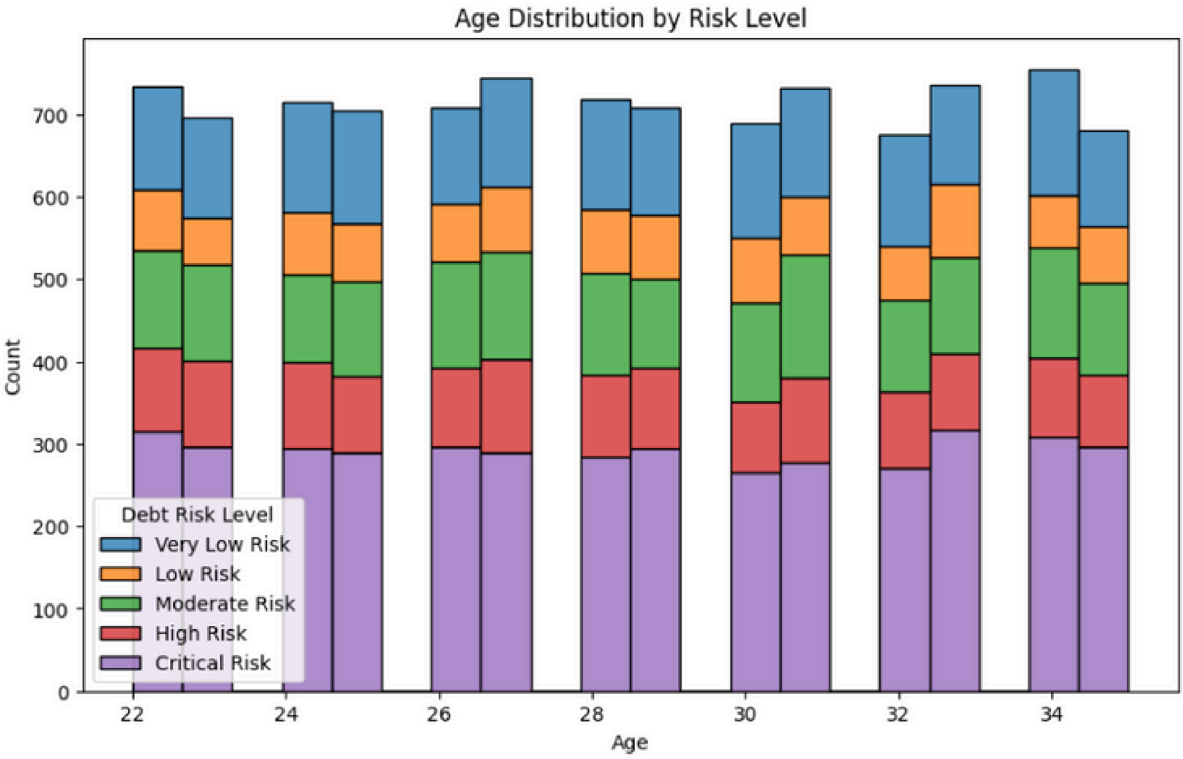
2.**Risk-Level Transitions:**

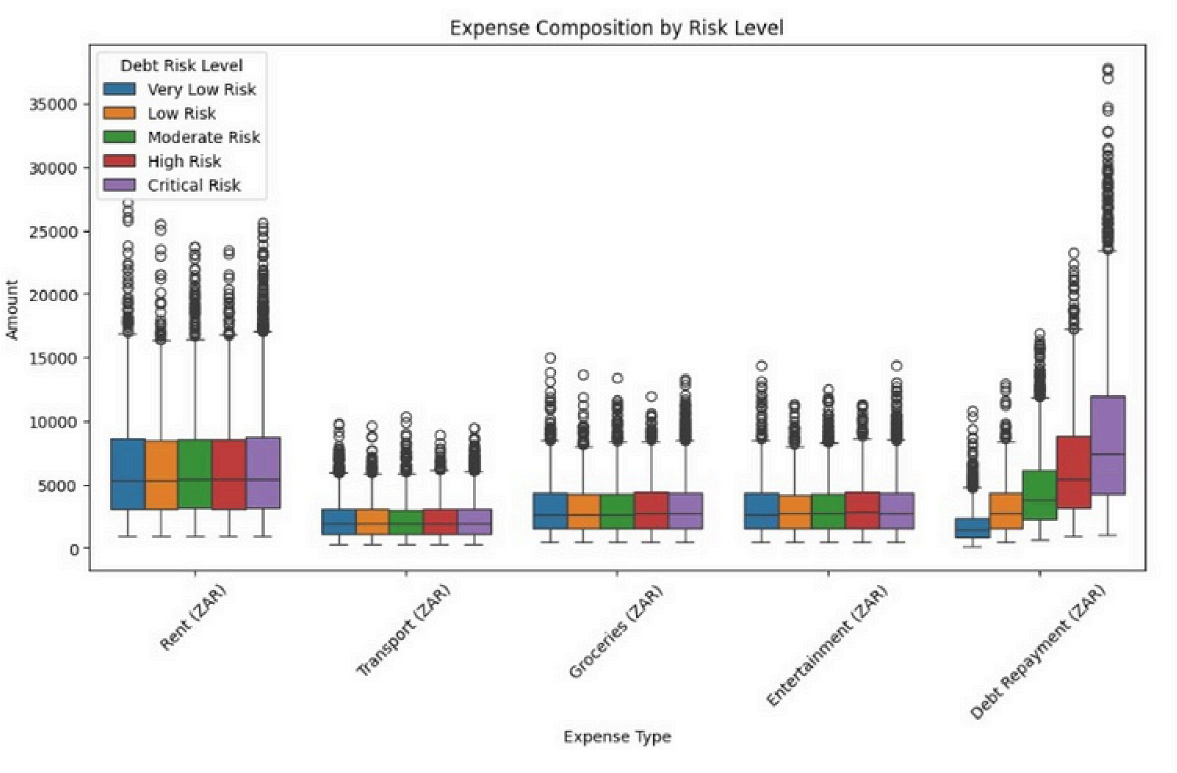
- **Critical → Moderate Risk** after 35 suggests "aging out" of worst debt stress.

o**Model Tip:** For 25–35-year-olds, prioritize **income growth potential** over current DTI.

3.**Outlier Groups:**

- **Older Critical Risk Borrowers (60+):** Rare but may signal emergencies (e.g., medical debt).





Title: Top 20 Feature Importance for Debt Risk Prediction

Description:

This horizontal bar chart ranks the most influential features in the XGBoost model by their relative importance (0–0.5 scale). Features are ordered from highest to lowest impact on predicting debt risk levels.

Key Findings:

1. Dominant Financial Ratios (Top 3):

- Expense_to_Salary_Ratio (0.48)

o***Why it matters:***The single strongest predictor—borrowers spending >50% of income on expenses are high-risk.

- Log_Salary (0.35)

o***Why it matters:***Higher income = lower risk, but logarithmic scaling shows diminishing returns.

- Salary_DTI_Interaction (0.32)

- ***Why it matters:*** Combines income and debt burden (e.g., high salary + high DTI = nuanced risk).

Binned Features (Moderate Impact):

- `Salary_Bin_Moderate` **(0.25)** & `Salary_Bin_High` **(0.22)**
  - ***Why it matters:*** Categorical salary ranges help the model segment risk nonlinearly.
- `Expenses_Bin_Moderate` **(0.15)**
  - ***Why it matters:*** Flags borrowers in ambiguous spending ranges.

3. Industry/Race Surprises (Low but Notable):

- `Industry_Healthcare` (0.08) & `Industry_Finance` (0.07)
  - ***Why it matters:*** Healthcare workers show slightly higher risk (likely due to variable roles), while finance workers are lower-risk.
- `Race_Indian/Asian` (0.06)
  - ***Proceed with caution:*** May reflect cultural savings habits, but avoid over-reliance to prevent bias.

4. Unexpected Low Importance:

- `Savings (ZAR)` (0.05)
  - ***Why?*** Likely overshadowed by expense/salary ratios—savings matter less if income covers debts.
- Demographics (`Gender`, `Race_Black African`)
  - ***Not in top 20:*** Confirms the model prioritizes financial behavior over demographics.

- **Feature Engineering:**
  - **Create New Ratios:**

python

Copy

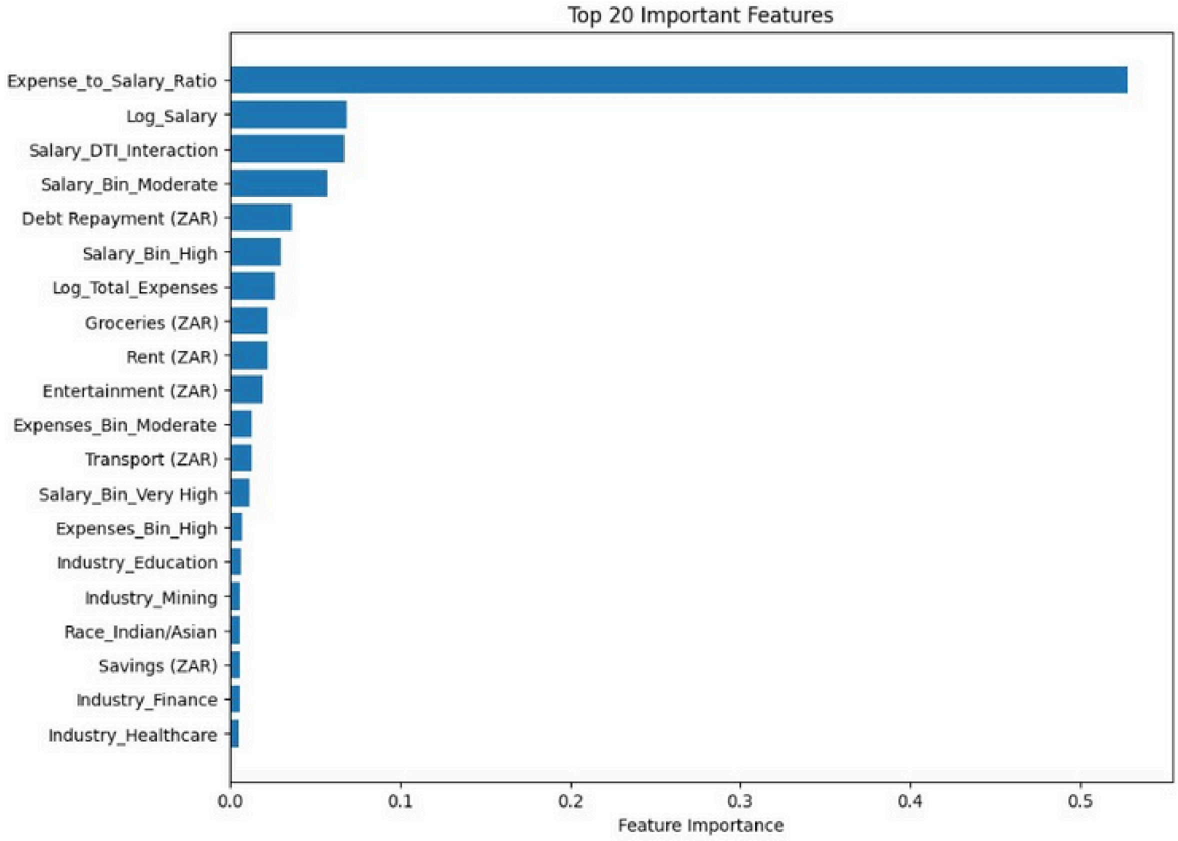
```
df['Essential_Spending_Ratio'] = (df['Rent (ZAR)'] + df['Groceries (ZAR)'])
/ df['Salary (ZAR)']
```

- ○ **Drop Low-Value Features:** Remove Transport (ZAR) and Entertainment (ZAR) (redundant with expense bins).

- **Interpretability:**
  - ○ **Explain to Stakeholders:**

    "Debt risk is 60% driven by three factors: spending ratios, salary, and DTI interactions—not demographics."

- **Bias Mitigation:**
  - ○ Audit predictions for Race_Indian/Asian to ensure no unfair advantages/disadvantages.


Top 20 Important Features
Expense_to_Salary_Ratio
Log_Salary
Salary_DTI_Interaction
Salary_Bin_Moderate
Debt Repayment (ZAR)
Salary_Bin_High
Log_Total_Expenses
Groceries (ZAR)
Rent (ZAR)
Entertainment (ZAR)
Expenses_Bin_Moderate
Transport (ZAR)
Salary_Bin_Very High
Expenses_Bin_High
Industry_Education
Industry_Mining
Race_Indian/Asian
Savings (ZAR)
Industry_Finance
Industry_Healthcare
Feature Importance