Here are some suggestions for features that could be collected via web crawling and merged with your existing credit risk dataset:

1. **Loan Intent-Specific Risk Trends**:
   - **Description**: Since `loan_intent` (e.g., PERSONAL, EDUCATION, MEDICAL) is available in your dataset, you can crawl data on default rates or risk profiles associated with different loan purposes. For instance, medical loans might have higher default rates due to unexpected health expenses.
   - **Source for Crawling**: Financial blogs, industry reports, or public datasets from financial institutions or government agencies (e.g., Consumer Financial Protection Bureau reports, or articles from sites like Investopedia or NerdWallet).
   - **How to Merge**: Create a mapping of `loan_intent` to average risk scores or default probabilities based on crawled data. For example, if reports indicate that MEDICAL loans have a 20% higher default rate on average, you can add a feature like `loan_intent_risk_factor` to your dataset with a numerical value representing relative risk for each category.
   - **Relevance**: This can provide context to the purpose of the loan and its inherent risk, enhancing the predictive power beyond just the category label.
2. **Home Ownership Status Risk Indicators**:
   - **Description**: The `person_home_ownership` feature (e.g., RENT, OWN, MORTGAGE) can be linked to general risk trends associated with housing status. For example, renters might have different financial stability profiles compared to mortgage holders.
   - **Source for Crawling**: Housing market reports, financial stability studies, or public data from real estate websites (e.g., Zillow Research, Redfin) or government housing statistics that discuss default rates or financial stress by ownership type.
   - **How to Merge**: Similar to loan intent, assign a risk factor or financial stress indicator to each category of `person_home_ownership` ba sed on crawled data. For instance, a feature like `ownership_risk_score` could be added with values reflecting relative risk (e.g., higher risk for RENT compared to OWN based on national trends).
   - **Relevance**: This adds depth to the home ownership feature by associating it with broader financial risk trends, which can influence creditworthiness.
3. **General Economic Indicators for Loan Grades**:
   - **Description**: Since `loan_grade` is in your dataset, you can crawl historical or general data on how different loan grades (A, B, C, D, etc.) correlate with default rates or economic conditions. While we don't have specific years, we can use long-term averages or trends that are not time-specific.
   - **Source for Crawling**: Financial institution reports, credit rating agency publications (e.g., Moody's, S&P Global), or public articles on loan grading systems from financial news websites.
   - **How to Merge**: Map `loan_grade` to a risk multiplier or default probability based on crawled data. For example, if grade D loans historically have a 30%

default rate in aggregated studies, add a feature like `loan_grade_risk` with corresponding values for each grade.
   - **Relevance**: This can contextualize the loan grade with empirical risk data, making the feature more informative for modeling.
4. **Industry-Wide Employment Length Norms**:
   - **Description**: While `person_emp_length` provides the duration of employment, it lacks context about what is typical or risky across industries or job types. Crawling data on average employment durations or turnover rates by sector can provide a benchmark to assess if a person's employment length is unusually short or long (potentially indicating instability or stability).
   - **Source for Crawling**: Labor statistics from government websites (e.g., Bureau of Labor Statistics in the US), job market analysis from sites like LinkedIn or Glassdoor, or industry reports on employment trends.
   - **How to Merge**: Since the dataset doesn't specify industry, you can use a general average across all sectors or focus on major sectors and apply a normalized risk factor. For instance, create a feature like `emp_length_risk` where values below the national average employment duration get a higher risk score. This would be a rough approximation but still useful.
   - **Relevance**: This helps interpret `person_emp_length` in a broader context, potentially identifying outliers who might be at higher risk due to short employment durations.
5. **Interest Rate Environment Context**:
   - **Description**: The `loan_int_rate` feature shows the interest rate for the loan, but without temporal data, we can still crawl long-term average interest rates or ranges for personal loans to see if a given rate is unusually high (indicating higher risk or subprime lending).
   - **Source for Crawling**: Historical interest rate data from financial websites, central bank reports (e.g., Federal Reserve), or consumer loan statistics from sites like Bankrate or LendingTree.
   - **How to Merge**: Add a feature like `int_rate_deviation` which measures how much the `loan_int_rate` deviates from the long-term average personal loan rate found via crawling. A higher-than-average rate could indicate a riskier borrower.
   - **Relevance**: This provides context to the interest rate, helping to differentiate between standard and high-risk loans based on market norms.

Considerations for Web Crawling:

- **Ethical and Legal Constraints**: Ensure that the data you crawl is publicly available and does not infringe on privacy or data protection laws. Focus on aggregated statistics rather than individual data.
- **Data Quality**: Verify the reliability of sources when crawling data. Government and reputable financial websites are generally more trustworthy.
- **Generalization**: Since we're applying broad trends to individual records without specific identifiers like time or location, the features will be approximations. However,

even rough contextual data can improve model performance by adding layers of interpretation to existing features.

Recommendation:

I suggest starting with **Loan Intent-Specific Risk Trends** and **Home Ownership Status Risk Indicators** as they can be directly mapped to categorical features in your dataset (`loan_intent` and `person_home_ownership`). These are likely to be the easiest to merge and can provide immediate value by contextualizing the purpose of the loan and housing status with risk data. You can crawl financial reports or consumer studies from reputable sources to gather this information.