# Project G2, author Sten Raak

# Predicting defaulting loans (within 12 months) using statistical and ML methods.

Repo: https://github.com/stenraak/Bondora_public_PD12

## Business Understanding

### Background:

Bondora, as a leading peer-to-peer lending platform, faces the ongoing challenge of minimizing loan defaults to ensure financial stability and customer satisfaction. This data-mining initiative aims to develop a robust predictive model that accurately forecasts whether a customer is likely to default on a loan within the first 12 months of borrowing. By leveraging the extensive public dataset provided by Bondora, the company aims to enhance risk assessment and make informed decisions to mitigate potential losses.

### Business Goals:

1. **Risk Mitigation:** Minimize financial losses by accurately identifying customers with a higher default risk.
2. **Customer Satisfaction:** Ensure a positive borrowing experience by approving loans for customers with lower default probabilities.
3. **Operational Efficiency:** Streamline loan approval processes through automation based on data-driven insights.
4. **Regulatory Compliance:** Align the lending practices with regulatory requirements, ensuring responsible lending.

### Business Success Criteria:

1. **Default Rate Reduction:** Achieve a measurable reduction in the overall default rate compared to historical data.

2. **Loan Approval Accuracy:** Improve the accuracy of loan approval decisions, leading to a decrease in false positives and false negatives.
3. **Customer Retention:** Maintain or increase customer satisfaction and loyalty by providing reliable risk assessments.

## Assessing Your Situation

### Inventory of Resources:

1. **Data:** Utilize the extensive Bondora public dataset, including customer demographics, loan details, repayment history, and other relevant features.
2. **Expertise:** Leverage the expertise of data scientists, statisticians, and domain experts to develop a comprehensive predictive model.
3. **Technology:** Utilize advanced analytics tools, machine learning algorithms, and computing resources for model development and deployment.

### Requirements, Assumptions, and Constraints:

1. **Data Quality:** Assumption that the Bondora dataset is accurate, complete, and representative of the lending environment.
2. **Regulatory Compliance:** Compliance with data privacy laws and lending regulations in all relevant jurisdictions.
3. **Resource Constraints:** Limited time and budget for model development and implementation.

### Risks and Contingencies:

1. **Model Accuracy:** Risk of inaccurate predictions leading to financial losses. Contingency: Continuous model monitoring and refinement based on new data.
2. **Regulatory Changes:** Risk of changes in lending regulations affecting the model's validity. Contingency: Regular updates and adjustments to align with evolving regulatory requirements.

### Terminology:

1. **Default:** Failure to repay a loan within the specified timeframe.
2. **False Positive:** Incorrectly predicting a customer will default when they do not.
3. **False Negative:** Incorrectly indicating a customer will not default when they do.
4. **Model Accuracy:** The ability of the predictive model to correctly identify default and non-default cases.

### Costs and Benefits:

1. **Costs:** Investment in technology, human resources, and time for model development, implementation, and ongoing maintenance.
2. **Benefits:** Reduction in default-related financial losses, improved customer satisfaction, and streamlined loan approval processes leading to operational efficiency.

## Defining Your Data-Mining Goals

### Data-Mining Goals:

1. **Predictive Model Development:** Develop an accurate machine learning model capable of predicting the likelihood of loan default within 12 months.
2. **Feature Selection:** Identify critical features influencing default and optimize the model by selecting the most relevant variables.
3. **Model Interpretability:** Ensure the model is interpretable, allowing stakeholders to understand and trust the predictions.

### Data-Mining Success Criteria:

1. **Model Accuracy:** Achieve high accuracy in predicting loan defaults, minimizing false positives and negatives.
2. **Interpretability:** Ensure the model provides clear explanations for its predictions, enhancing stakeholder trust.
3. **Feature Importance:** Identify and prioritize the most influential features affecting the default predictions.

By focusing on these specific goals, Bondora aims to harness the power of data mining to strengthen its risk management strategies, enhance customer experiences, and maintain a competitive edge in the dynamic landscape of peer-to-peer lending. Through diligent assessment and continuous improvement, the company endeavors to achieve long-term success and sustainability in its lending practices.

# Data understanding

### Gathering Data

*Outline Data Requirements:*

In the context of predicting loan defaults within 12 months based on Bondora's public dataset, the following data requirements have been identified:

1. **Historical Loan Data:** Information on past loans, including details on borrowers, loan amounts, interest rates, and loan terms.
2. **Borrower Demographics:** Data on borrower characteristics such as age, employment status, income level, and geographical location.
3. **Credit Score Information:** Scores that reflect the creditworthiness of borrowers, helping assess the risk of default.
4. **Income and liabilities assessment:** Consistent repayments and stable income can suggest a potential good customer.
5. **Loan Performance Metrics:** Metrics indicating the success or failure of loans, including whether a borrower defaulted within the first 12 months.

*Verify Data Availability:*

The Bondora public dataset is verified to contain information relevant to the outlined data requirements. However, it's crucial to acknowledge the limitations due to certain features being populated post-loan issuance. This restricts the use of those features as predictive variables in the model.

Also note that since some data is protected by regulations, then we do not actually have all the data that Bondora does to predict if the customer will default.

### *Define Selection Criteria:*

The selection criteria involve choosing features that are available pre-loan issuance and are both predictive and interpretable for default prediction. Features such as credit scores, borrower demographics, historical loan performance, and employment stability are prioritized for inclusion.

## Describing Data:

### *Exploring Data:*

Exploration of the Bondora public dataset reveals a diverse set of features, including but not limited to:

1. **Loan Details:** Loan amount, interest rate, and term.
2. **Borrower Information:** Age, employment status, income level, and geographical location.
3. **Credit Scores:** If available, these scores play a vital role in predicting creditworthiness.
4. **Repayment History:** Insights into previous repayment behavior, helping anticipate future patterns.

The exploration involves:

- Examining the distribution of numeric variables.
- Identifying categorical variables.
- Assessing potential relationships between features.

Visualization tools can create histograms, scatter plots, and correlation matrices to understand the data better.

### *Verifying Data Quality:*

Data quality is crucial for model development. Steps taken to verify data quality include:

1. **Missing Values:** Identify and address any missing values in critical features. Impute or remove missing data based on the extent and impact of the model.
2. **Outliers:** Evaluate the presence of outliers that might skew predictions. Consider whether outliers are genuine data points or errors that require

correction.

3. **Data Consistency:** Ensure consistency in data representation and units. For example, standardize income values to a common currency.
4. **Data Accuracy:** Cross-reference data with external sources or conduct internal audits to verify the accuracy of critical information.
5. **Incorrect data:** Make sure there are no odd values that do not match with the general idea of the data.

Documenting data quality checks and any transformations made to enhance the dataset's reliability is essential.

### Conclusion:

The data understanding phase of the CRISP-DM process for the Bondora loan default prediction project has been comprehensive. I was gathering data involved outlining specific requirements, verifying availability, and defining selection criteria to address the challenges posed by post-issuance features. Describing data encompassed exploring the dataset, identifying key features, and visualizing relationships between variables. Verify data quality focused on addressing missing value outliers and ensuring data consistency and accuracy.

This robust data understanding lays the foundation for subsequent phases in the project, providing a clear roadmap for feature selection, model development, and continuous improvement. As the project progresses, ongoing data exploration and quality monitoring will be essential to adapt to changing patterns and maintain the effectiveness of the predictive model.

## Project Plan (with approximate time estimations)

### 1. Data Preprocessing and Cleaning (15 hours):

- *Tasks:* Handle missing values, address outliers, standardize data formats, and perform any necessary transformations.
- *Methods:* If the majority of data exists, impute its data based on the KNN model otherwise, remove the features, scaling numerical variables with Standard- or RobustScaler (helps with outliers)

### 2. Feature Selection and Engineering (10 hours):

- *Tasks:* Identify relevant features, explore interactions, and create new variables for improved model performance.

### 3. Model Development and Evaluation (15 hours):

- *Tasks:* Select appropriate machine learning algorithms, split data into training and validation sets, train models, and evaluate performance metrics.
- *Methods:* Utilize scikit-learn for model implementation, employ cross-validation for robust evaluation, and use optuna for hyperparameter optimization.

### 4. Documentation and Presentation (10 hours):

- *Tasks:* Document the entire process, create a comprehensive model report, and prepare a presentation for stakeholders.
- *Tools:* Use Jupyter Notebooks for documentation to create visually appealing presentations with tools like Matplotlib and Seaborn.

### Important Notes:

- Utilize version control systems (e.g., Git) to track changes and ensure reproducibility.
- Python will be the primary programming language, leveraging libraries such as scikit-learn, pandas, and numpy for data manipulation and analysis.