# Prosper Loan Portfolio Valuation: Project Overview & Comprehensive Plan

This document provides a comprehensive overview of the "Prosper Loan Portfolio Valuation" project, detailing what has been accomplished, the current state of the database, and a clear, granular roadmap for completing all remaining phases. This information should enable a new team member to seamlessly pick up where the project left off and serve as the definitive project plan for all future work.

## 1. Project Summary & Overall Goal

**Overall Project Goal**: To value defined segments of a historical personal loan portfolio from the prosperLoanData.csv dataset using a segmented Discounted Cash Flow (DCF) model, incorporating industry benchmarks and sensitivity analysis, with a primary focus on robust data preparation and SQL-based data processing.

## 2. Key Guiding Decisions & Parameters

* **Dataset**: prosperLoanData.csv (Prosper loans originated ~2005-2014).
* **Primary Focus Portfolio**: 'Core Non-Performing' loans (LoanStatus = 'Chargedoff' or 'Defaulted').
* **Secondary Analysis (Optional)**: 'At-Risk' loans (LoanStatus = 'Past Due 90-120 days', 'Past Due >120 days').
* **Valuation Date**: January 1, 2015 (to align with data vintage).
* **Forecast Horizon**: 84 months from the valuation date.
* **Score Proxy**: Midpoint of CreditScoreRangeLower and CreditScoreRangeUpper. ProsperScore is used with flagging/imputation for missing values, and potentially as a separate segmentation variable.
* **Recovery Base**: LP\_GrossPrincipalLoss for the 'Core Non-Performing' portfolio.
* **Key Model Assumptions (Stored in model\_parameters table)**:
  + valuation\_date: '2015-01-01'
  + forecast\_horizon: '84' months
  + base\_case\_discount\_rate: '0.20' (20%)
  + base\_case\_collection\_cost\_percentage: '0.25' (25%)

## 3. Accomplished Phases & Key Milestones So Far

We have successfully completed the foundational data engineering and database setup, which is the cornerstone for the DCF model. This includes:

* **Database Design**: Designed and created the debt\_portfolio PostgreSQL database with a structured schema, including:
  + staging\_prosper: For raw data ingestion.
  + portfolio\_accounts\_core: For cleaned, filtered, and feature-engineered 'Core Non-Performing' loan data.
  + score\_risk\_mapping: A lookup table for FICO score bands and risk multipliers.
  + recovery\_curve\_parameters: A lookup table for segment-specific monthly recovery rates.
  + model\_parameters: A configuration table for global DCF assumptions.
* **Robust Data Cleaning & Feature Engineering**:
  + Successfully loaded raw data into staging\_prosper.
  + Developed and iteratively refined complex SQL INSERT statements to populate portfolio\_accounts\_core. This involved extensive data cleaning, type casting (handling problematic text, empty strings, and decimals to NUMERIC/INT/DATE), and filtering for 'Chargedoff' and 'Defaulted' loans.
  + Engineered crucial features: charge\_off\_date, fico\_score\_proxy, loan\_age\_at\_charge\_off\_months, months\_since\_charge\_off\_at\_valuation.
  + Crucially, a robust fix for the prosper\_score column was implemented, ensuring it accurately reflects original scores (e.g., 4, 6) instead of 0 by accommodating decimal points in the source data, with missing/non-numeric values now correctly defaulting to NULL.
  + estimated\_loss and estimated\_return were added to portfolio\_accounts\_core, leveraging Prosper's own risk metrics for future analysis, with NULL handling for missing values.
* **Segmentation Logic**: The initial 3x4 segmentation (FICO bands x Loan Original Amount quartiles) was implemented to assign a segment\_id to each loan in portfolio\_accounts\_core.
* **Lookup Table Population**: score\_risk\_mapping was successfully populated with defined risk multipliers, and recovery\_curve\_parameters with 84 months of segment-specific recovery percentages, applying base rates and risk adjustments.
* **Global Parameter Setup**: model\_parameters was populated with key DCF assumptions (Valuation Date: Jan 1, 2015; Forecast Horizon: 84 months; Base Discount Rate: 20%; Collection Cost: 25%).
* **Initial Portfolio Metrics**: total\_gross\_principal\_loss (approx. $79.81M), forecasted\_gross\_recovery (approx. $4.66M), and gross\_recovery\_rate (approx. 5.84%) were calculated, establishing baseline figures.
* **Current Status**: The database is fully structured and populated with clean, prepared, and segmented data. The foundational SQL scripts for data preparation are complete and verified. The database is now ready to serve as the direct input for the DCF model.

### 4. Database Schema (debt\_portfolio)

The debt\_portfolio PostgreSQL database has the following structured schema:

* **staging\_prosper**: Raw imported data.
  + **Columns**: All original columns as per prosperLoanData.csv (all TEXT type).
* **portfolio\_accounts\_core**: Cleaned, filtered 'Core Non-Performing' loans with engineered features.
  + **Columns**: listing\_key (PK), loan\_key, loan\_original\_amount (NUMERIC), lp\_gross\_principal\_loss (NUMERIC), loan\_origination\_date (DATE), closed\_date (DATE), loan\_status (VARCHAR), charge\_off\_date (DATE), fico\_score\_proxy (INT), loan\_age\_at\_charge\_off\_months (INT), months\_since\_charge\_off\_at\_valuation (INT), prosper\_score (INT, nullable), estimated\_loss (NUMERIC, nullable), estimated\_return (NUMERIC, nullable), segment\_id (VARCHAR).
* **score\_risk\_mapping**: FICO proxy band to risk multiplier lookup table.
  + **Columns**: score\_band (PK, VARCHAR), risk\_multiplier (NUMERIC).
* **recovery\_curve\_parameters**: Segment-specific monthly recovery percentages lookup table.
  + **Columns**: segment\_id (PK, part 1, VARCHAR), month\_number (PK, part 2, INT), recovery\_percent (NUMERIC).
* **model\_parameters**: Global model assumptions configuration table.
  + **Columns**: parameter\_name (PK, VARCHAR), parameter\_value (VARCHAR), description (TEXT), unit (VARCHAR), last\_updated (DATE).

### 5. Important Notes for Data Handling & Current State

* **Snake\_case Naming**: All database tables and columns are consistently named using snake\_case.
* **Robust Casting**: All TEXT to NUMERIC/INT/DATE conversions from staging\_prosper are handled robustly using TRIM, NULLIF, COALESCE, CAST, and CASE statements to manage empty strings and non-numeric values.
* **prosper\_score**: This column now correctly reflects original values (0-11 range based on prosperdataset.docx, but actual data observed 1-10) after the fix for decimal parsing and proper NULL handling for missing/non-numeric values.
* **estimated\_loss & estimated\_return**: These are now correctly available in portfolio\_accounts\_core with NULLs for missing data.
* **Recovery Rate Benchmark**: The current calculated gross recovery rate of ~5.84% is lower than industry benchmarks for charged-off unsecured personal loans (typically 10-30%). This suggests a potentially conservative recovery curve or warrants further investigation/sensitivity analysis in later phases.
* **Loan Seasoning for Recoveries**: The absence of forecasted gross recoveries for forecast months 1-9 is expected and accurate, as all loans contributing significant principal losses in the portfolio\_accounts\_core are seasoned by at least 9 months at the January 1, 2015 valuation date.

## 6. Comprehensive Roadmap for Project Phases

The project has completed phases up to **Phase V: Sensitivity & Scenario Analysis Setup**. This roadmap details remaining steps.

### Phase I: Project Setup & Foundational Research (Completed)

* **Step 1: Formalize Objectives & Scope**
  + 1.1: Write a final, concise objective statement for the 'Core Non-Performing' portfolio valuation.
  + 1.2: Briefly outline the objective for the (optional) 'At-Risk' portfolio analysis (e.g., "To estimate potential future charge-offs and their indicative value").
  + 1.3: List key deliverables (e.g., Valuation Model, Final Report, Presentation, SQL Code Repository).
* **Step 2: Establish Valuation Parameters**
  + 2.1: Confirm and document the Valuation Date: January 1, 2015 and Forecast Horizon: 84 months.
* **Step 3: Gather & Refine Recovery Benchmarks**
  + 3.1: Document the EBA benchmarks (Net Recovery Rate: ~33-38%, Time: ~3 years) as a starting point.
  + 3.2: Actively research and document any available US-specific recovery data for unsecured personal loans (from ~2010-2018 period if possible) to contextualize or adjust EBA data.
  + 3.3: Note limitations of applying EU benchmarks to US data and how US context will be used.
* **Step 4: Develop Discount Rate Methodology**
  + 4.1: Research and select a primary methodology for determining the discount rate.
  + 4.2: Identify components (Risk-Free Rate, Market Risk Premium, Illiquidity Premium, Specific NPL Risk Premium).
  + 4.3: Establish a justifiable discount rate range and a base case rate.
* **Step 5: Define Score Proxy & Mapping Strategy**
  + 5.1: Confirm use of CreditScoreRangeLower/Upper midpoint as primary FICO proxy.
  + 5.2: Define bands for this proxy (<620, 620-679, 680-719, 720+).
  + 5.3: Outline how ProsperScore (1-11) will be handled (flagging missing, imputation using SQL, potential separate segmentation).
  + 5.4: Plan how these score bands will be used to adjust base recovery benchmarks (e.g., applying risk multipliers).
* **Step 6: Outline Collection Cost Assumption**
  + 6.1: Select a base case flat percentage for collection costs (e.g., 20-30% of gross recovered amounts).
  + 6.2: Plan for sensitivity analysis.

### Phase II: Data Acquisition, Cleaning & Initial Exploration (SQL-Focused) (Completed)

* **Step 7: Load & Audit Data using SQL**
  + 7.1: Load prosperLoanData.csv into a SQL database staging table.
  + 7.2: Perform initial data audit using SQL queries (counts, schema, nulls, distinct values).
  + 7.3: Review data types and identify cleaning needs.
* **Step 8: Execute Detailed Data Cleaning using SQL**
  + 8.1-8.7: Handle missingness, validate fields, standardize values, parse dates, convert monetary fields, handle duplicates, detect/handle outliers using SQL.
* **Step 9: Filter for Portfolios & Engineer Features using SQL**
  + 9.1: Create the 'Core Non-Performing' portfolio table/view (portfolio\_accounts\_core).
  + 9.2: (For later) Create the 'At-Risk' portfolio table/view (portfolio\_accounts\_at\_risk).
  + 9.3: For 'Core' portfolio, engineer features (Charge\_Off\_Date, FICO\_Score\_Proxy, Loan\_Age\_at\_Charge\_Off\_Months, Months\_Since\_Charge\_Off\_at\_Valuation).
* **Step 10: Initial Exploratory Data Analysis (EDA) - Core Portfolio using SQL**
  + 10.1: Generate descriptive statistics for key variables.
  + 10.2: Generate data for visualizations (histograms, box plots) using SQL group by queries.
  + 10.3: Investigate initial hypotheses using SQL queries.
* **Step 11: Create/Update Data Dictionary**
  + 11.1: Document all original fields and newly engineered features.

### Phase III: Database Setup & Advanced EDA (Completed)

* **Step 12: Setup SQL Database** (Covered in Step 7.1)
* **Step 13: Design & Create Normalized Tables using SQL**
  + 13.1: portfolio\_accounts\_core (created in Step 9.1).
  + 13.2: Create Recovery\_Curve\_Parameters table in SQL.
  + 13.3: Create Score\_Risk\_Mapping table in SQL.
* **Step 14: Load Data & Perform SQL-based EDA** (Covered in Steps 7-10)
  + 14.1: Load cleaned 'Core' portfolio data into portfolio\_accounts\_core. Populate helper tables (Recovery\_Curve\_Parameters, Score\_Risk\_Mapping).
  + 14.2: Practice SQL EDA.

### Phase IV: Core Non-Performing Portfolio Valuation Model (Completed - Base Model)

* **Step 15: Finalize Segmentation Strategy using SQL**
  + 15.1: Implement initial 3x4 segmentation using SQL.
  + 15.2: Verify segment sizes.
* **Step 16: Implement Recovery Curve Application using SQL**
  + 16.1: Select/adapt a base monthly recovery curve. Store this in Recovery\_Curve\_Parameters.
  + 16.2: Apply risk multipliers by joining portfolio\_accounts\_core with score\_risk\_mapping and then with recovery\_curve\_parameters.
  + 16.3: Crucial: For each loan, determine its starting month on its segment's recovery curve based on Months\_Since\_Charge\_Off\_at\_Valuation.
  + 16.4: Unit Test with a small set of loans.
* **Step 17: Develop DCF Model (Excel or a dedicated financial modeling tool, fed by SQL-prepared data)**
  + 17.1: Export data from SQL (LP\_GrossPrincipalLoss, segment, seasoning, recovery curve parameters).
  + 17.2: Aggregate monthly gross recoveries for each segment (partially prepared using SQL SUM() and GROUP BY).
  + 17.3: Apply collection costs to gross recoveries to get net monthly cash flows per segment (in the DCF tool).
  + 17.4: Discount net monthly cash flows for each segment back to the Valuation Date.
  + 17.5: Aggregate segment NPVs to get total portfolio NPV.

### Phase V: Sensitivity & Scenario Analysis (Core Portfolio) (Completed - Setup & Calculation)

* **Step 18: Identify & Model Key Sensitivities**
  + 18.1: Focus on: Recovery Rate Curves (proportional shift), Discount Rate, Collection Costs.
* **Step 19: Perform Sensitivity Analysis**
  + 19.1: Implement one-way data tables for each key variable vs. Portfolio NPV.
  + 19.2: Implement two-way data tables (e.g., Discount Rate vs. Recovery Multiplier).
* **Step 20: Define & Analyze Scenarios**
  + 20.1: Define Base, Upside, and Downside scenarios.
  + 20.2: Calculate Portfolio NPV for each scenario.
* **Step 21: Visualize Results**
  + 21.1: Create a Tornado chart to show the impact of key sensitivities.
  + 21.2: Chart base case projected cash flows and cumulative recoveries.

### Phase VI: At-Risk Portfolio Analysis (Secondary Analysis - Optional)

* **Goal**: Estimate potential future charge-offs and their indicative value from the 'At-Risk' portfolio.
* **Step 22: Analyze 'At-Risk' Portfolio Characteristics using SQL**
  + 22.1: Perform EDA on the portfolio\_accounts\_at\_risk table/view using SQL queries.
* **Step 23: Model Transition Probabilities**
  + 23.1: Research typical roll rates/transition probabilities for delinquent personal loans.
  + 23.2: Apply these probabilities to the 'At-Risk' portfolio segments using SQL queries.
* **Step 24: Estimate Value of Potential Future Charge-offs**
  + 24.1: For the estimated future charge-offs, apply the valuation framework from Phase IV (using appropriate timing assumptions).

### Phase VII: Deliverables, Reporting & Final Polish

* **Goal**: Finalize all project outputs and prepare for presentation.
* **Step 25: Develop Written Report & Presentation**
  + 25.1: Draft Executive Summary.
  + 25.2: Write Full Report/Slide Deck detailing: Introduction, Data Sources & Preparation (SQL-based), Valuation Methodology, Assumptions, Results, Sensitivity, Limitations, Conclusion.
* **Step 26: Organize & Finalize Code & Models**
  + 26.1: Thoroughly comment all SQL scripts. Ensure they are runnable and well-organized.
  + 26.2: If using Excel for the DCF model, ensure it is clean, well-labeled, with formulas protected and a README sheet.
* **Step 27: Conduct Peer Review**
  + 27.1: Have a peer or mentor review your report, model logic, and SQL code for clarity, correctness, and completeness.
* **Step 28: Create Project Repository**
  + 28.1: Set up a GitHub repository. Include SQL code, a summarized version of the report, the cleaned dataset (if permissible), and a README.md explaining the project and how to run/interpret it.

**Contact**: LadrillonoJustin@Gmail.com

Please reach out if you have any questions or require further clarification on the completed work. Good luck with the remaining phases!