Goal:

Build predictive models that explain and forecast loan outcomes using real-world financial data, and interpret what these models reveal about borrowing and credit markets.

Data: ***Links are in Google Colab, I loaded them first thing.***

Prosper set: Official Prosper Loan Data, which includes detailed information on hundreds of thousands of personal loans

HMDA: Home Mortgage Disclosure Act, a federal dataset published by the CFPB (Consumer Financial Protection Bureau)

- contains detailed information about almost every mortgage application in the U.S.

- Introduction: Overview of the data, predictive task, and summary findings.
- Data Description: Data source and description
- Models and Methods: Overview of models and implementation
- Results and Interpretation: Review of modeling results and interpretation of performance
- Conclusion and Next Steps: Summary of models and next steps for further analysis

1. How do interest rates vary by loan purpose (debt consolidation, auto, home improvement, etc.)?

Model used: Linear Regression

Why: Simple, interpretable model that shows how categorical loan purposes affect APR.

Findings:

- Loan purpose has very small predictive power relative to other features.
- APR differences across purposes are minimal once controlling for credit-related variables.
- Lenders do not price interest rates based on loan purpose; credit risk dominates.

2. Are certain borrower demographics or financial characteristics (DTI, credit score, employment length) associated with higher/lower loan amounts?

Model used: Artificial Neural Network (MLPRegressor)

Why: Captures nonlinear patterns between income, DTI, employment status, credit metrics, and loan size.

Findings:

- The ANN achieved moderate predictive power ($R^2 \approx 0.53$).
- Income and credit rating variables are the strongest predictors of loan amount.

- Employment status and categorical variables contribute but with much smaller magnitudes.

3. How does loan grade (A–G) relate to both interest rate and risk metrics?

Model used Artificial Neural Network (MLPRegressor)

Why: Captures nonlinear patterns between income, DTI, employment status, credit metrics, and loan size.

Findings:

- The ANN achieved moderate predictive power ($R^2 \approx 0.53$).
- Income and credit rating variables are the strongest predictors of loan amount.
- Employment status and categorical variables contribute but with much smaller magnitudes.

4. Do borrowers with similar income levels cluster into distinct loan-amount groups?

Model used gradient Boosting Regressor

Why: Powerful nonlinear model ideal for ranking feature importance and capturing complex pricing rules.

Findings:

- Prosper Rating (Alpha) is overwhelmingly the strongest predictor of APR.
- Ratings E, HR, and D sharply increase APR; AA sharply decreases it.
- Loan purpose, income, and employment have minimal impact on APR once credit grade is included.
- The model explained ~93% of variance in APR → confirms strong underwriter-driven pricing.

5. Are there nonlinear patterns in how credit score affects interest rate (e.g., diminishing returns past a certain score)?

Model: Polynomial Regression (Degree 2)

Why: Captures curvature and diminishing returns in credit-score effects.

Findings:

- APR decreases as credit score increases, but not linearly.
- The largest APR reductions occur from 600 → 720.
- Above ~750, APR declines flatten → diminishing marginal benefit of excellent credit.
- The polynomial model slightly outperformed linear, confirming nonlinear behavior in terms of r2 score.

Introduction

Advances in data availability and machine learning have transformed how consumer credit markets evaluate risk and determine loan pricing. Rather than relying solely on relationship based lending or simple credit rules, modern platforms increasingly use automated underwriting systems that incorporate borrower characteristics, credit quality, and loan attributes to set interest rates and loan sizes. Understanding how these variables interact is important both for lenders seeking to manage risk and for borrowers navigating credit markets.

This project studies the determinants of interest rates and loan amounts using data from two sources. The first is the Prosper peer to peer lending platform, which provides detailed information on borrower credit characteristics, loan purposes, and pricing outcomes. The second is the Home Mortgage Disclosure Act dataset, which offers insight into income and loan size patterns in the mortgage market. By combining supervised learning and unsupervised clustering techniques, this analysis explores how credit risk is priced, how borrower characteristics affect loan size, and whether borrowers naturally segment into distinct groups.

The central hypothesis of the project is that borrower credit risk variables dominate loan pricing decisions, while loan purpose and demographic characteristics play a more limited role once risk is accounted for. A secondary hypothesis is that the relationship between credit score and interest rates is nonlinear, with diminishing improvements at higher score levels.

Data Description

Two datasets are used in this analysis.

The Prosper dataset contains information on consumer loans issued through a peer to peer lending platform. Key variables include interest rates, loan amounts, borrower income, credit score, debt to income ratio, employment length, and loan purpose. This dataset is used to estimate regression based models for both interest rates and loan amounts.

The HMDA dataset contains mortgage level observations for New York State. The primary variables used are borrower income and loan amount. This dataset is used exclusively for clustering analysis in order to identify whether borrowers form distinct income and loan size groups.

Models and Methods

Different modeling approaches are chosen based on the structure of each research question.

Linear regression is used when the goal is to estimate average relationships and maintain interpretability, particularly for assessing how loan purpose relates to interest rates.

A multilayer perceptron regressor is used to predict loan amounts because borrower characteristics may interact in complex and nonlinear ways that are not well captured by linear models.

Gradient boosting regression is used to model interest rates with high predictive accuracy and to capture nonlinear interactions among borrower risk characteristics.

K means clustering is used to identify borrower income and loan amount groupings in an unsupervised setting where no outcome variable is specified.

Polynomial regression is used to test whether the relationship between credit score and interest rates exhibits nonlinear behavior.

All models are implemented using scikit learn, and performance is evaluated using out of sample R squared where applicable.

Across the supervised learning tasks, model performance varies in informative ways that reflect the underlying economic mechanisms of lending. For interest rate prediction, simpler linear models perform adequately for interpretability but leave substantial unexplained variation, while more flexible models capture the structure of pricing decisions far more effectively. This contrast highlights the extent to which modern credit pricing relies on nonlinear combinations of borrower characteristics rather than additive effects alone. The strong performance of nonlinear models in predicting APR suggests that pricing rules are systematic and algorithmic, rather than discretionary.

In contrast, predictive performance for loan amount is more moderate. Even with flexible models such as neural networks, a significant share of variation remains unexplained. This gap is consistent with the idea that loan size decisions depend not only on observable borrower characteristics but also on platform-specific constraints, borrower self-selection, and policy rules that are not fully encoded in the data. As a result, loan amounts appear to be shaped by a mix of measurable repayment capacity and institutional factors that are harder to observe directly.

The unsupervised clustering results complement these findings by showing that borrowers naturally group into discrete income and loan size tiers. Rather than exhibiting smooth continuous variation, the data suggest segmentation consistent with underwriting thresholds and affordability bands. Taken together, these results reinforce the conclusion that credit markets

operate through structured risk classification systems that govern both pricing and allocation decisions.

Results and Interpretation

The first question was whether interest rates vary meaningfully by loan purpose (debt consolidation, auto, home improvement, and so on). The linear regression results suggest that loan purpose provides very little predictive power once we control for credit-related variables. In practice, APR differences across purposes shrink to small values after accounting for borrower risk factors. The interpretation is that lenders are not using "purpose" as a major pricing input; instead, the interest rate is primarily a function of credit risk. Loan purpose may look like it matters in raw averages, but once you control for the borrower's credit profile, it stops being a strong driver of pricing.

The second question asked whether borrower demographics and financial characteristics, DTI, credit score, employment length, are associated with higher or lower loan amounts. For this, we used the MLPRegressor, and the model achieved moderate predictive power with $R^2 \approx 0.53$. This is meaningful but not "near-perfect," which makes sense because loan amount decisions can include lender policies, borrower preferences, and constraints that aren't fully captured in the dataset. The strongest signals came from income and credit rating variables, which fits the logic of underwriting: borrowers with higher repayment capacity and stronger credit profiles are typically approved for larger amounts. Employment status and other categorical features do contribute, but their effect is noticeably smaller compared to income and credit quality. The overall takeaway is that loan size is shaped most strongly by ability-to-pay and creditworthiness rather than demographic categories by themselves.

The third question examined how loan grade relates to interest rates. This relationship is consistent with the overall pattern in the project: interest rate pricing is dominated by credit risk classification. Once loan grade and related credit measures are included, other variables such as loan purpose and employment characteristics add relatively little incremental explanatory power. The key interpretation is that loan grade functions as a compressed underwriting signal, and interest rates largely follow that risk ranking.

The fourth question asked whether borrowers with similar income levels cluster into distinct loan amount groups. This part of the analysis used K-means clustering on the HMDA dataset with borrower income and loan amount as inputs. The results show distinct groupings that resemble affordability tiers: a lower income and lower loan cluster, a middle income and moderate loan cluster, and a higher income and higher loan cluster. Because this is an unsupervised method, the output is interpreted as segmentation patterns rather than predictive accuracy, and no R^2 is reported.

The fifth question focused on whether the credit score–APR relationship shows diminishing returns. The polynomial regression (degree 2) confirms that APR decreases as credit score rises, but not at a constant rate. The steepest improvement happens as borrowers move from roughly 600 to 720, where APR declines most strongly. Above around 750, the curve flattens and APR reductions become much smaller, indicating that once you are already considered very low risk, additional score gains do not change pricing much. In terms of model fit, the polynomial specification slightly outperformed the linear model on R^2, which supports the claim that the relationship is nonlinear rather than purely linear.

Conclusion and Next Steps

Putting everything together, the project shows a consistent pattern: credit risk dominates lending outcomes, especially interest rate pricing. Loan purpose and many demographic or categorical variables may appear meaningful at first glance, but once we include strong credit indicators like loan grade and Prosper Rating (Alpha), their independent impact becomes small. The modeling also highlights that pricing is not only risk-driven but also nonlinear, especially around the credit score range where borrowers transition from "moderate risk" to "low risk." For loan amounts, predictive performance is moderate rather than extremely high, which suggests that loan size decisions depend on a mix of measurable borrower capacity and other factors not fully captured in the dataset.

For next steps, there are several natural extensions. One direction would be to make the clustering question more literal by applying explicit clustering methods (like k-means or hierarchical clustering) to loan amount patterns conditional on income rather than treating it mainly through a supervised model. Another direction would be deeper model interpretation: for the gradient boosting model, tools like partial dependence or SHAP-style explanations could help show exactly how Prosper Rating (Alpha) and credit variables shift APR across the distribution. Finally, it would be useful to validate whether these patterns hold consistently across time or subsets of borrowers, which would test how stable the underwriting rules are.

From a repository and reporting standpoint, the cleanest structure is to separate exploration from modeling and then use those notebooks to support a standalone written report. A typical setup would include one notebook dedicated to EDA (variable distributions, missingness, correlations, key plots) and a separate notebook for modeling (preprocessing, model training, evaluation, and

interpretation). The final written report could summarize the motivation, methods, results, and conclusions in one place, while the notebooks provide the reproducible evidence behind the claims.