

# Peer to Peer Lending Investor Dashboard

Team 14: Clayton Francis, Nicholas Park, Robson De Souza, Bingqian Hu, Xiaolu Su  
Georgia Tech Computer Science & Engineering, Data and Visual Analytics

## Motivation

P2P lending market is an under-invested market and investors often fail to effectively diversify

The Peer to Peer (P2P) lending market is a decentralized platform where individuals borrow and lend with one another.

Investors are often under-informed, either by funding loans at random or over-pursuing excessively high returns which often leads to default.

Growth in P2P lending would have a substantial social impact where borrowers get the financing they need, and lenders can expect a positive return. A key to growing this market is to further educate investors.

Lenders need tools to guide well-informed investment decisions

Current platforms offer APIs allowing access to their data. Using historical loan data, we model the probability of default for a set of test loans.

Innovation of our project is showcased in an interactive dashboard where investors are informed of market analytics and provided with a pre-built list of optimized diversified portfolios.

## Data

Publicly available, raw large data size

The Lending Club API has both historical and current loans, but the data we used was downloaded from the Kaggle repository and only contains loans issued from 2007 to 2020.

**Data excluded from modeling:** loans before 2015, 60-month term loans, "Current" loans (still in repayment), and columns (features) which were intuitively insignificant to our analysis

Test Set	Most recent 10,000 loans
Validation Set	Next most recent 10,000 loans
Training Set	Remaining data

Key characteristics about both the borrower (# of credit inquiries, current revolving credit balance, etc.) and the issued loan (interest rate, term length, etc.).

700.0 K Number of applications  
\$12,000.0 Median Loan Amount  
10,432.59 M Total Loan Amount (market size)

- Size of raw data: 1.65 G, which include 700,000 records
- Features include numerical and categorical variables

## Conclusions

P2P Lending Market is a sufficient alternative investment that deserves greater attention and funding

Investors need access to validated statistical analysis which accounts for data features that humans cannot.

Designed pipeline for future use to process data on new loans and predict the probability of default more accurately than any generic risk profile assigned by platforms. Identifying significant borrower and loan characteristics provides additional insight to lenders.

With our easy-to-use lending tool, investors can now quickly gain a deep understanding about key market factors and have access to portfolio suggestions designed to minimize the probability of default while maximizing return, likely outperforming the market (according to experiment).

## Approaches

## Portfolio Assistant

Interactive dashboard designed to reduce emotional investing and over-funding risky loans by suggesting portfolios which have been modeled to maximize return while minimizing risk for a given risk profile and level of diversification.

To build optimal portfolios we need to maximize the expected return for each portfolio which is limited to a unique set of constraints, solved by PuLP (optimization software).

$$ExpectedReturn = [(1 + i[x]) \times (1 - DP[x])]$$

Where i = interest rate, DP = default probability (predicted by the credit risk model).

High interest loans are discounted by the predicted default probability. The resulting portfolios are loaded into the dashboard which displays the underlying loans to fund, the expected return, cash flows, Sharpe ratio, and resulting level of diversification.

Presenting investors with preset options reduces uncertainty and emotional investing, educates about power of diversification.

Growth in P2P lending dependent on widespread acceptance driven by well-informed investors consistently rewarded with positive returns.

## Market Insights

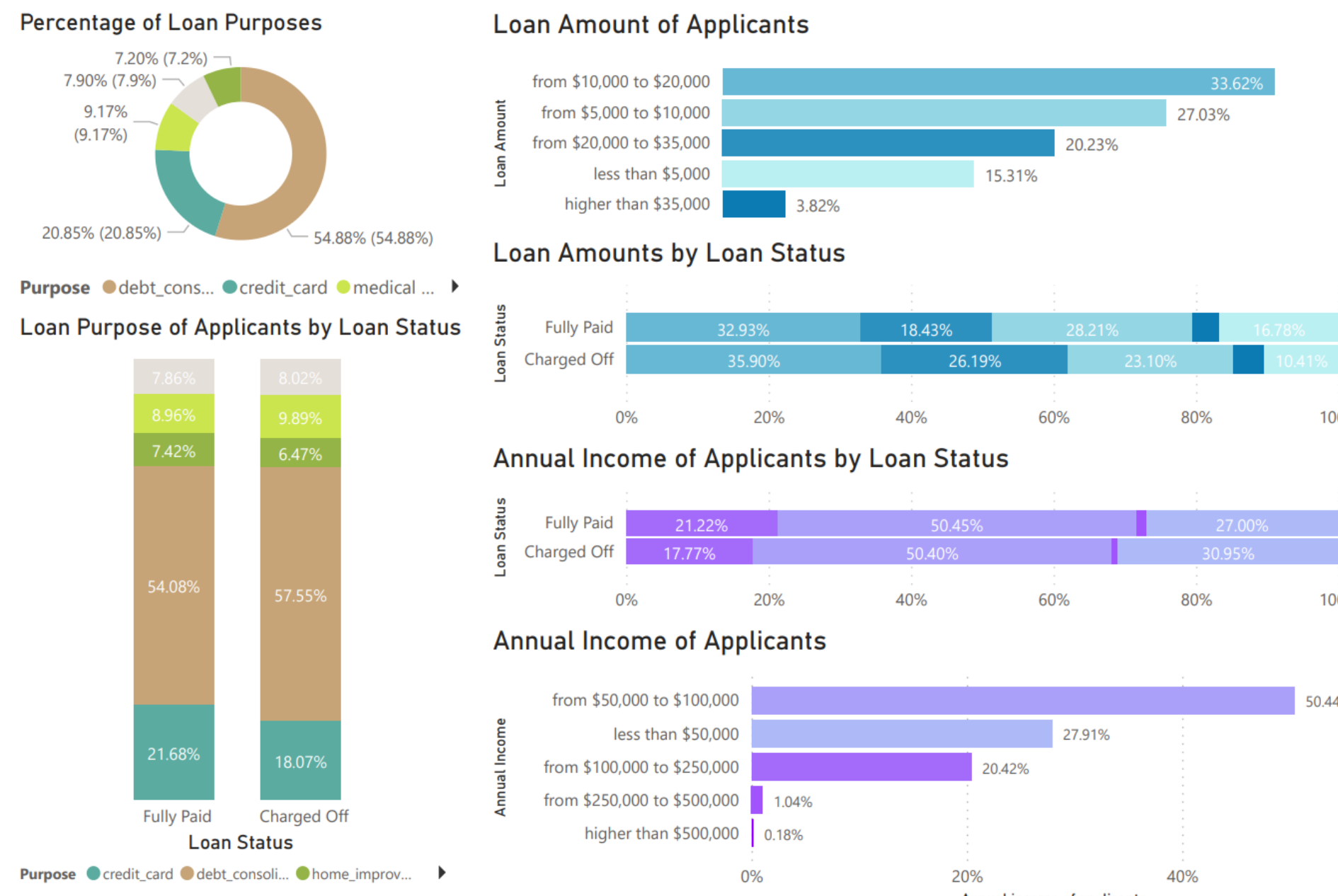
### Overview of P2P market containing...

- Summary statistics of loan and borrower data characteristics
- Feature importance visuals
- Trend charts of how loan performance differs across key sub-groups of model inputs

- P2P borrowers do not post collateral, which is unnatural to the average fixed income investor and is a systematic risk that increases the chances of losing entire investment

- Well-informed lenders are more likely to make recurring investments

- Descriptive statistics and understanding the drivers of default can be extremely beneficial to new market participant's ability to assess risk



## Credit Risk Model

**Input:** Borrower's credit bureau variables, demographic features  
**Output:** Probability of borrower defaulting on loan

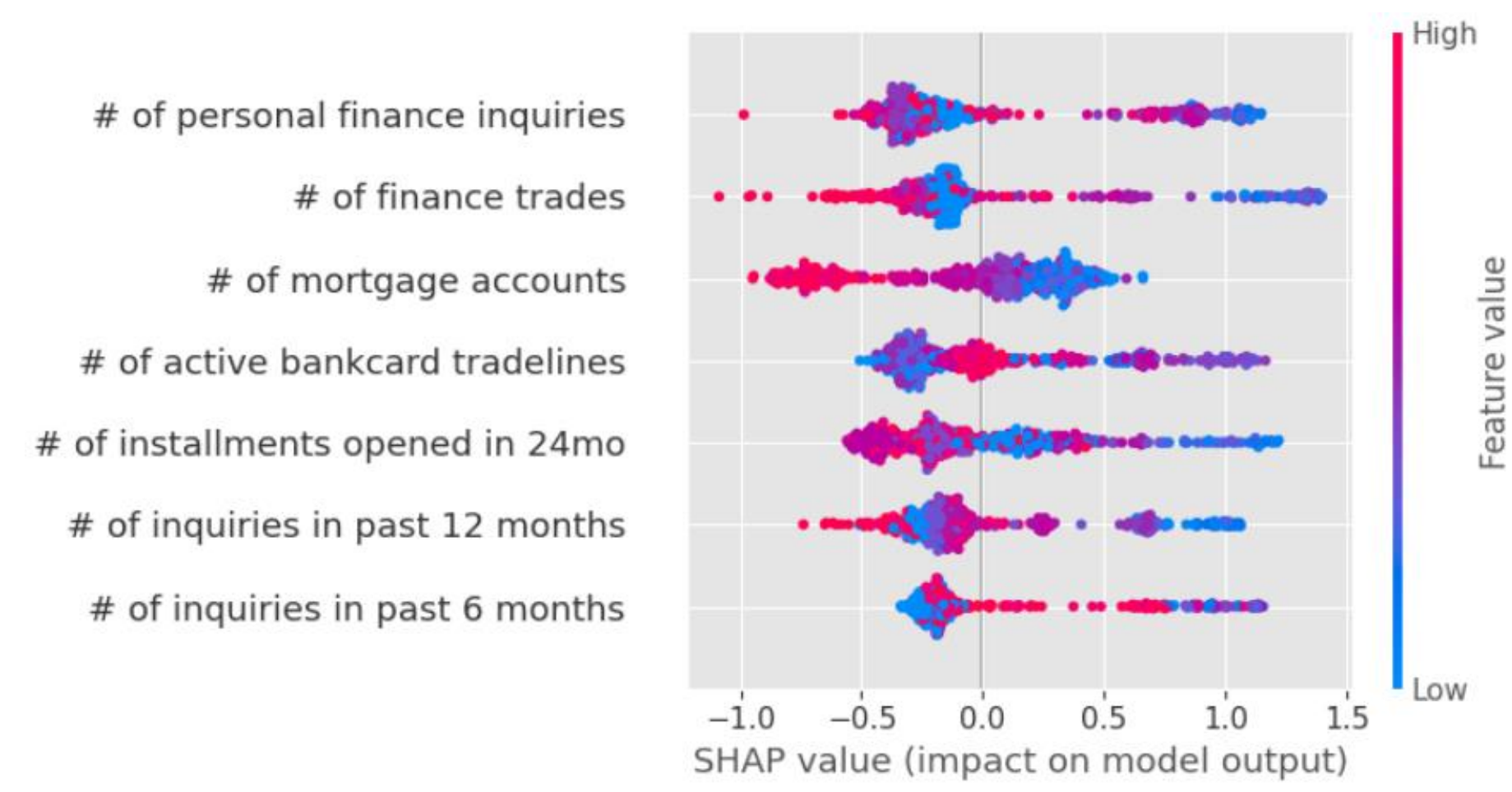
Developed a model class in Scikit-Learn that is called using a config file to specify hyperparameters and other configurable rules. Model pipeline entails...

- Feature preprocessing
  - Under/oversampling to handle heavy class imbalance
  - Lasso feature selection
  - ML model fitting
- \*Jupyter notebooks used to train multiple models and evaluate accuracy\*

Concluded that a stacked ensemble of CatBoost models yielded the highest accuracy score, outperforming peer classifiers such as logistic regression and neural networks.

P2P platform's proprietary loan-grading gives investors a relative idea of risk, but we take this a step further by observing the outcomes of historical P2P loans and modeling a deeper understanding of the driving factors of default.

Evaluated feature impact using the SHAP (SHapley Additive exPlanations) library, a visualization based on game theory to help explain ML model outputs.



## Experiments

### Optimized Portfolio Experiment

-> How do we know our "optimized" portfolios perform well?

Calculated performance and compared it to a benchmark. Benchmark portfolios consist of 5 random loans from the test set. Each loan has an interest rate +/- 3% the average interest rate of the comparable optimized portfolio. Random portfolios designed to impersonate uninformed investors lacking diversification.

	Optimized	Random
Average Return	3.75%	-3.07%
Average Months to Outcome	2.54	2.54
# of Defaulted Loans	0	18
Average Annualized Return	18.35%	-7.45%
Best Portfolio (Annualized)	38.0%	30.0%
Worst Portfolio (Annualized)	8.0%	-100.00%
Outperforming Portfolios	44	16
Sharpe Ratio	2.7	-0.17

#### Winner: Optimized Portfolio

- Higher returns
- Less underlying loan defaults
- Less volatility of returns

Expected time horizon is a few short months, good for reinvestment.

Attractive investment compared to other fixed income or equity products.

### Credit Risk Model Experiment

-> How does our model's predictions compare to Lending Club's proprietary loan grading model?

Compared our determination of risk to Lending Club's across the test set of loans. Our model's prediction for default probability and Lending Club's assigned interest rate had a correlation coefficient of **0.43, which defines a moderately positive relationship**.

Binned loans into groups based on different ranges of default probability to get a correlation plot that reduces noise and better visualizes the relationship, supporting the conclusion that our model is relevant and effective.

