SERVICE ANALYTICS

Prosper Replication Project

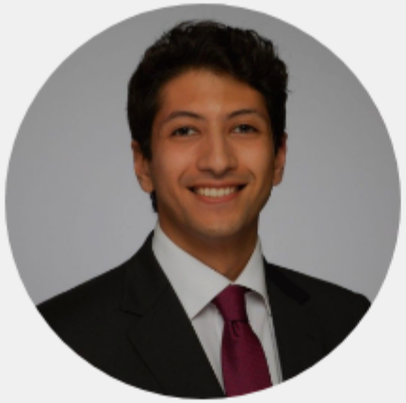
Our Team

Final Project

Shaan Kohli

Data Scientist

in



Ramy Hammam

Data Analyst

in

Shaher Yar Jahangir

Project Manager

in

Noah Mukhtar

Financial Analyst

in

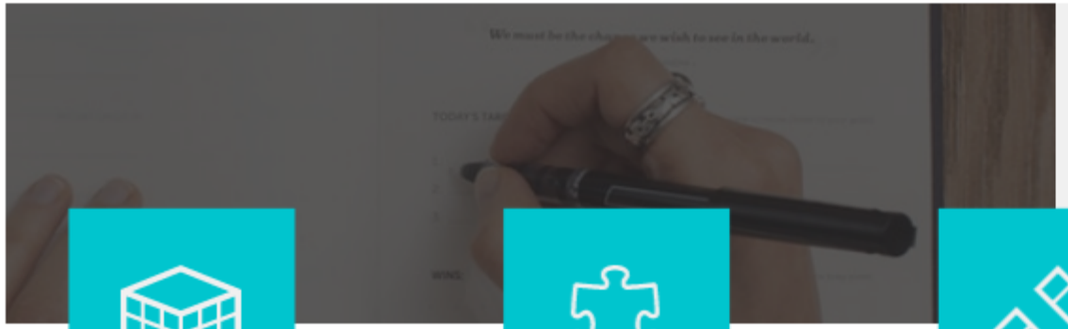
Replication Project

Choice: Option 2

Replicate LendingClub analysis using an alternative P2P platform - Prosper

Strategy: Predict which loans are more likely to default, and predict the amount value of the returns

Agenda



Data Preprocessing

Exploratory Data Analysis

Classification Model

Evaluation

Future Extensions

Intro. & Objectives

Part 1

Dataset

Powered by:

- Prosper Loan Data API

Time Period:

- 2014-2017

3 Excel files

- Each representing a year of gathered loan data

Loan Prediction

0: Completed

1: Default

Objective

Create a model that lets us invest in a portfolio made up of successful loans.

Dataset Comparison

Prosper vs. LendingClub

PROSPER

Features	Number
Predictors	22
Time Series	Yearly Data
Possible Outcomes	4
Outcome Types	Completed, Charged Off, Current, Defaulted
Size	×4 smaller than LendingClub

LendingClub

Features	Number
Predictors	+100
Time Series	Quarterly Data
Possible Outcomes	7
Outcome Types	Completed, Charged Off, Current, Defaulted
Size	x4 bigger than Prosper

Prosper Data Description

Updating

- Partially Updated Annually
- Partially Held Constant

Data Leakage

- High correlation with target variable





Categorization

- Integer
- Float / Decimal
- Categorical / String
- Binary (T/F)
- Date

Total Payment / Return

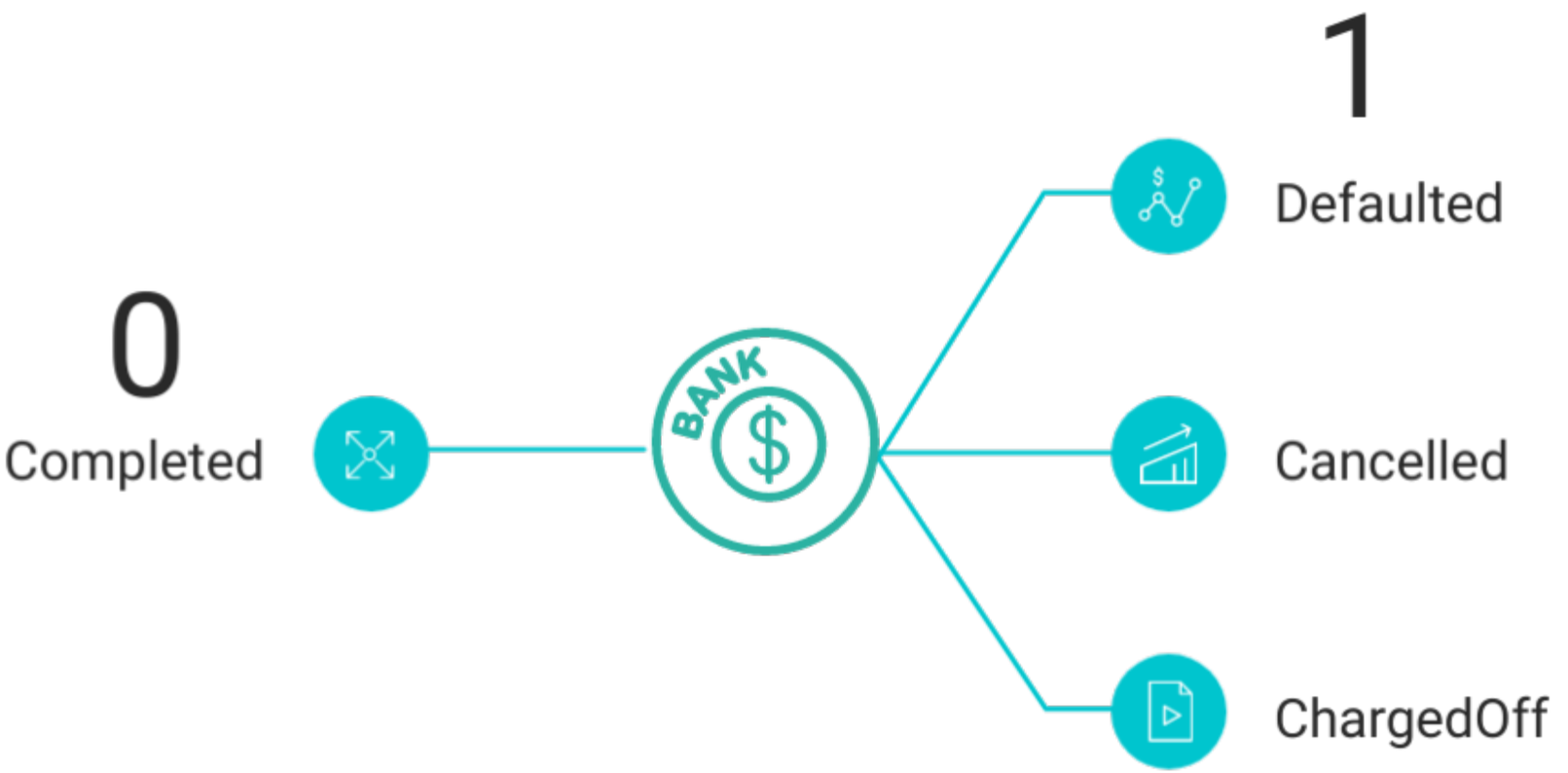
- Lack of such variable
- Needs to be calculated

Data Leakage

- Unavailability of predictor information during prediction of total return

Loan Status

Problem: Imbalanced Dataset



Treating Information Leakage

Avoid using highly correlated predictors:

- loan_number
- late_fees
- age_in_months
- days_past_due
- origination_date
- principal_balance
- principal_paid
- interest_paid
- late_fees_paid
- debt_sale_proceeds_received
- loan_default_reason
- loan_default_reason_description
- next_payment_due_date
- next_payment_due_amount
- co_borrower_application

15 Predictors

High correlation with (y)

Unavailable When Predicting Whether a Loan Will Default or Not

Time Leakage

- Value of prosper rating can be updated overtime
- i.e.: fico_range_low from LendingClub
- eg: loan defaulting causing borrower's prosper_score to decrease in the future
- Sets a pattern of high correlation between target variable and prosper_rating

Our Investor's Decision Making Process

Quantity To Invest

- How much money to invest (potential revenue)
- How much to allocate to other options for investment

Types of Loans

- Decide which loans to invest budget into
- Central focus of our loan classification model

Define a "Good" Loan

Vague definition means that the objective should be refined to whether:

1) Early Default of a Loan
2) Early Payback of a Loan
3) Time of Default / Time Taken to Payback Early

Data Ingestion & Cleaning

Part 2

Data Ingestion & Cleaning

Import Packages

- Numpy
- Pandas
- Matplotlib
- Seaborn

Merging Data: "df"

- 2014
- 2015
- 2016

Binary Transformation

0: Defaulted

1: Completed

Term Selection

- Subsetting according to LendingClub
- Filtering for greater than 36 months

Manipulating Variables

- Eliminating data leakage predictors & missing values
- Rename variables
- Creating new variables
  1) Total Payment
  2) Loan Length

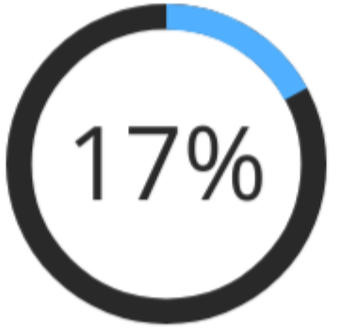
Correlation Check

Top 10 Positive Correlations

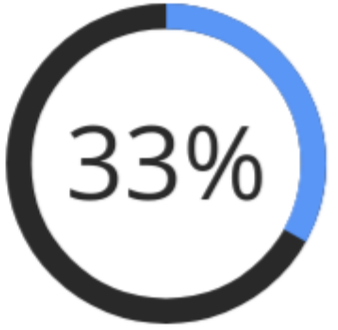
Term

Service Fees Paid

Borrower Rate



Days Past Due

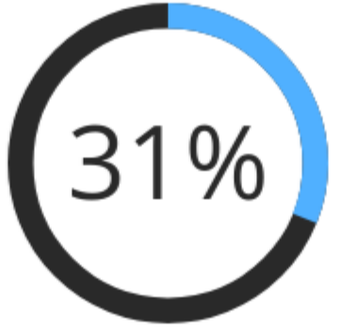


Debt Sale Proceeds Received

Late Fees Paid

Age in Months

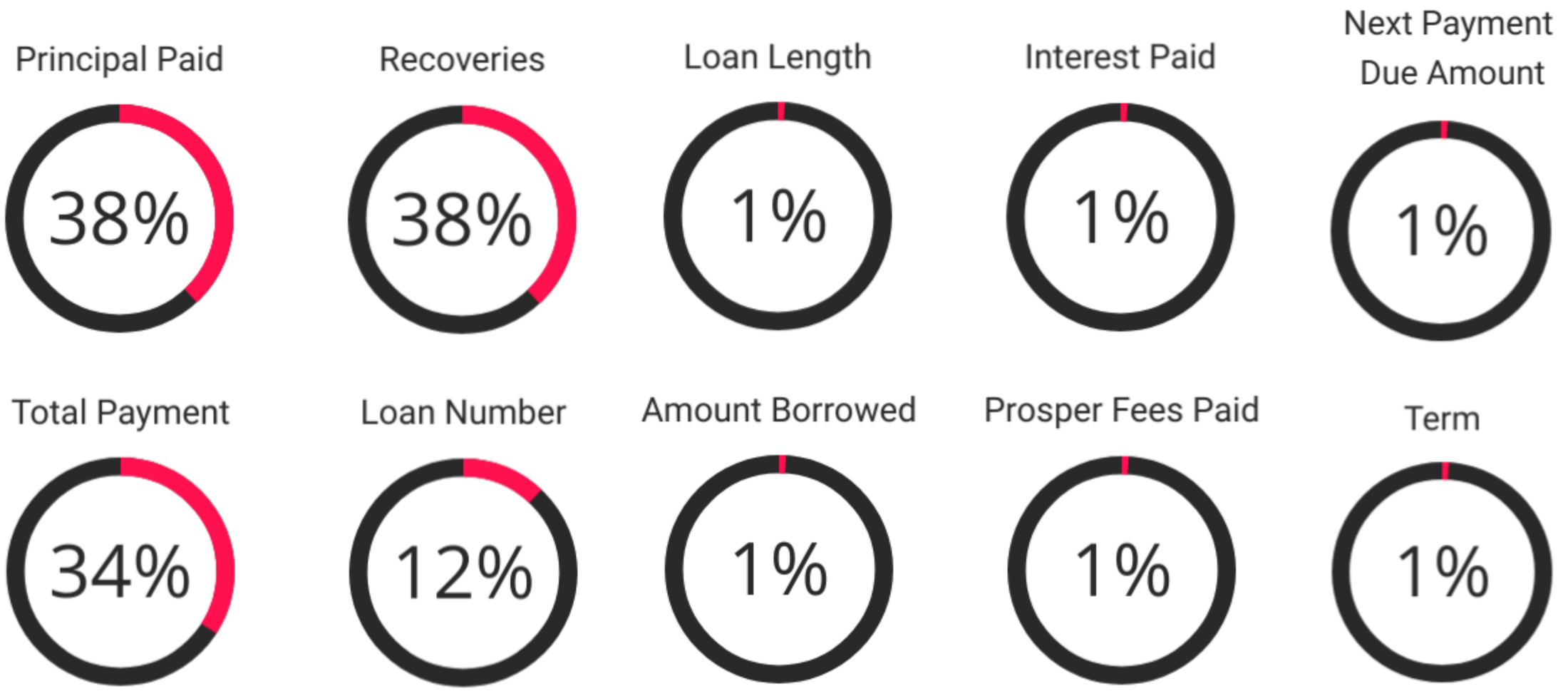
Principal Balance



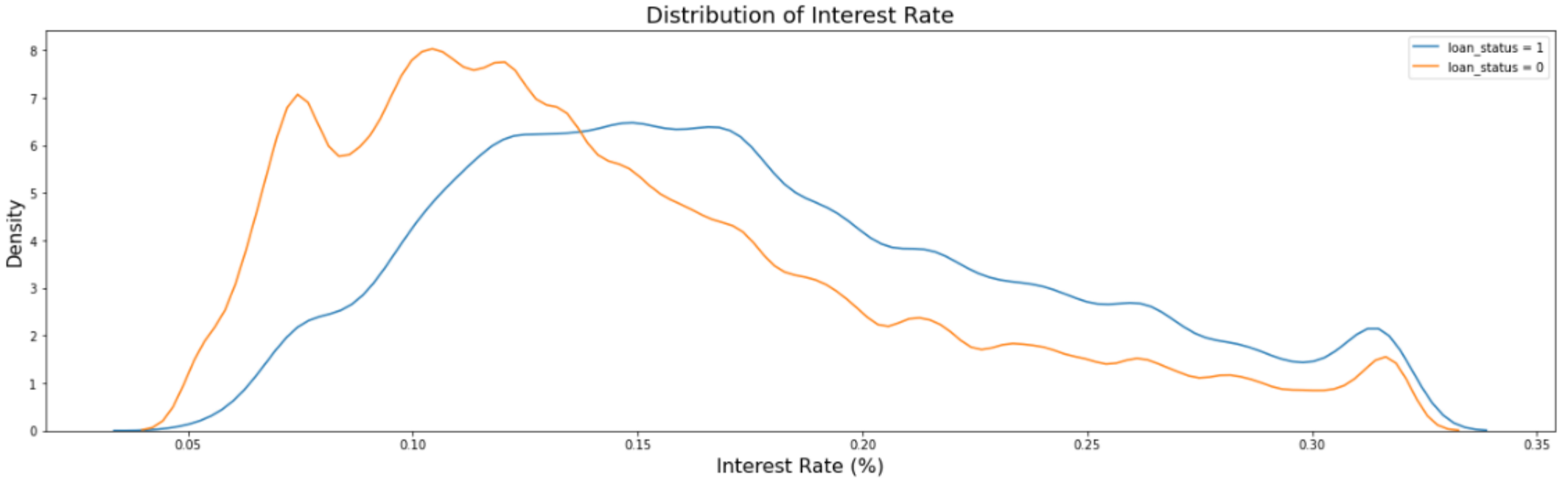
Co-Borrower App.

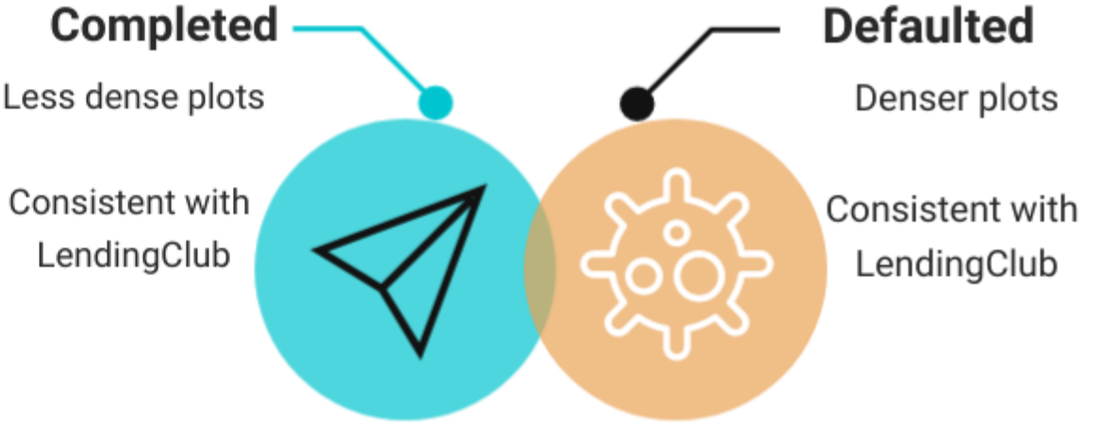
Correlation Check

Top 10 Negative Correlations



Distribution of Interest Rates





Loan Amounts Per Term

36 months

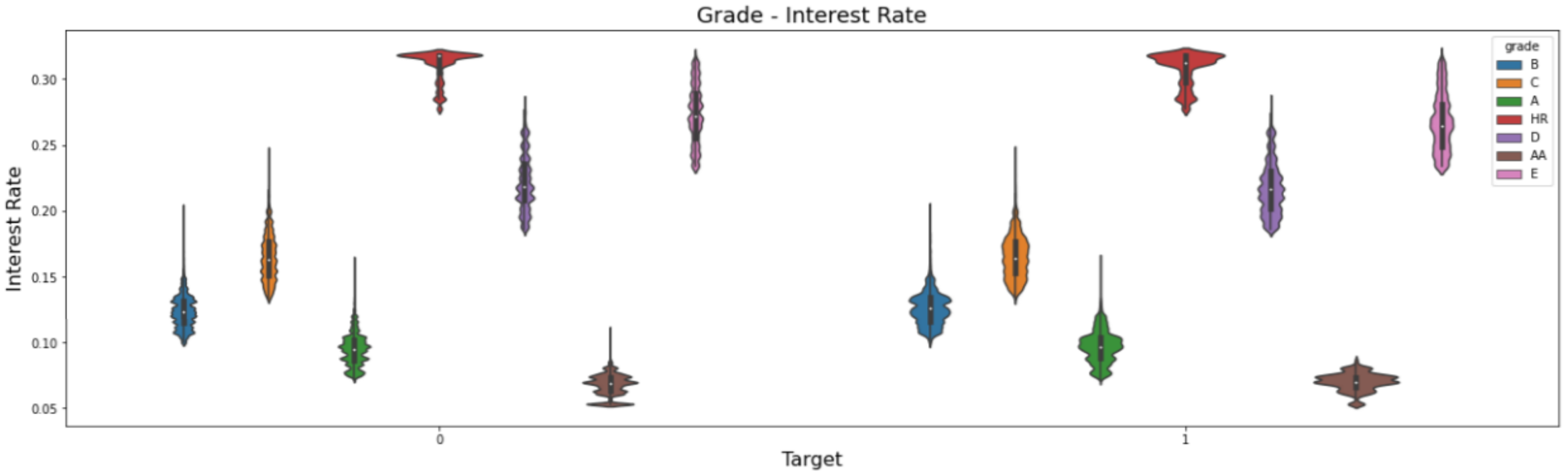
Much smaller amounts

More consistently distributed for both completed & defaulted

60 months

Highest: $15,000

Interest Rate Distribution Per Grade



Data Exploration

Part 3

Creating Our Variables

- Most significant data in determining potential return: **total payments**
- Building an investment strategy means picking a strong indicator variable of return on each loan
- Replicate LendingClub in terms of return considering partially paid off defaulted loans & early paid loans

Total Payment

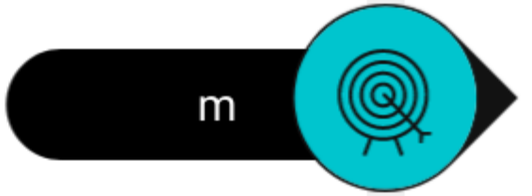
Principal Paid Amount + Interest Paid

Total Amount Invested in Loan

i.e.: Amount Borrowed By Bororwer

Term of Loan

i.e.: Nominal length of loan in months



Loan Length

i.e.: Actual length of loan in months

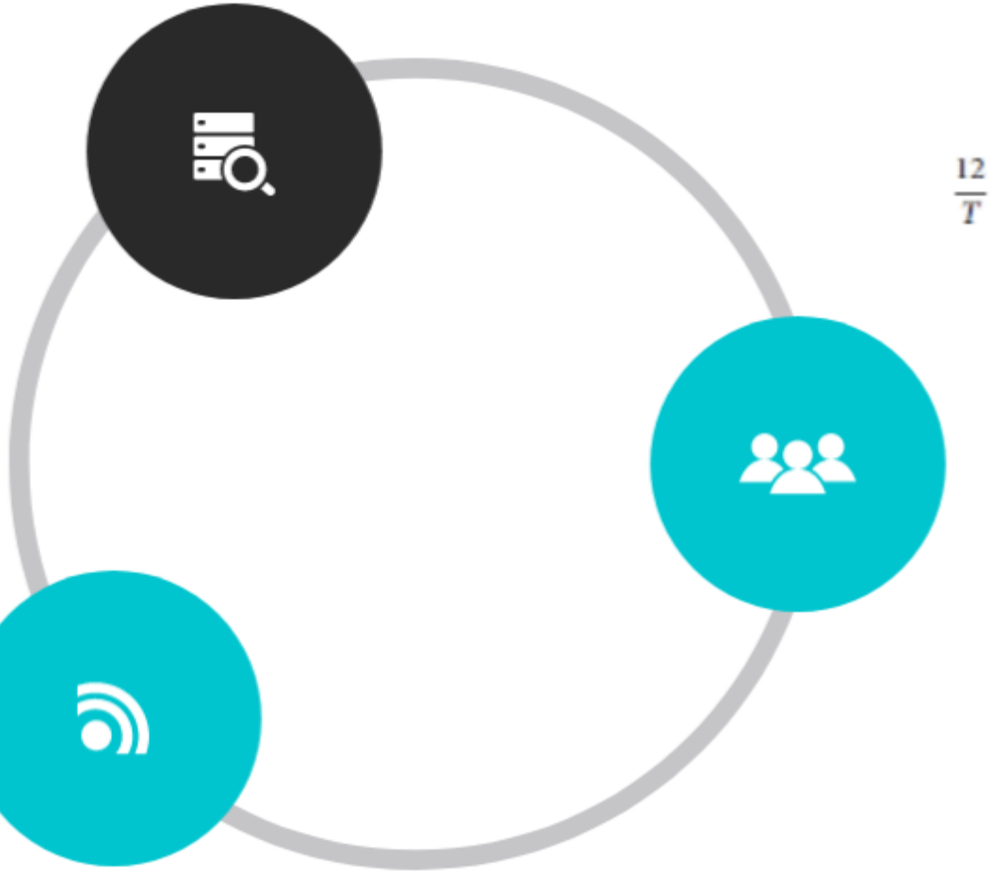
The 3 Measures

Pessimistic: M1

Loan is paid back & investor cannot re-invest until the term of the loan

$$\frac{p-f}{f} \times \frac{12}{t}.$$

Favours long-term loans



Optimistic: M2

Loan is paid back; investor can immediately re-invest in another loan with the same return

$$\frac{p-f}{f} \times \frac{12}{m}.$$

Assumption may result in huge overestimates of negative returns

Short & long loans are treated equally with this approach

$$\frac{12}{T} \cdot \frac{1}{f}\left\{\left[\frac{p}{m}\left(\frac{1-(1+i)^m}{1-(1+i)}\right)\right](1+i)^{T-m} - f\right\}.$$

FTHI: M3

3 Interest Rates: 1%, 2.5%, 5%

Equation is replicated

Most accurate methodology as it equalizes differences between loan length & defaulted loans

Disregards depreciation time of the value of money

Prosper Score Breakdown

Similarities

- Deeper dive into "grade" variable to replicate summary stats & explore trends for each return (M1-M3)
- Prosper shows similar trend to LendingClub
- Higher the grade, lower the interest & default rates

Grade	% of loans	% Default	Av. interest	Mean return			
				M1	M2	M3 (1.2%)	M3 (3%)
A	16.68	6.33	7.22	1.66	3.89	2.05	3.71
B	28.86	13.48	10.85	1.58	5.01	2.02	3.68
C	27.99	22.41	14.07	0.62	5.39	1.39	3.02
D	15.44	30.37	17.54	0.05	5.71	0.92	2.51
E	7.59	38.83	20.73	−0.91	5.95	0.11	1.64
F	2.72	44.99	24.47	−1.43	6.43	−0.44	1.05
G	0.73	48.16	27.12	−2.58	6.66	−1.40	0.05

Key differences:

- Prosper stats indicate returns are very identical with minor increases across different grades
- No negative return values across any of the grades

	grade	default	int_rate	return_PESS	return_OPT	return_INTa	return_INTb	return_INTc
A	19.973938	8.612483	9.474686	2.274137	4.541028	2.254222	3.486214	5.706550
AA	10.377824	4.014627	6.775790	2.019641	3.706281	2.023254	3.275717	5.534159
B	22.979313	12.956079	12.394934	2.395357	5.703969	2.466247	3.724494	5.997924
C	24.597005	18.515979	16.501650	2.495612	7.242789	2.583270	3.886307	6.246494
D	12.239538	23.292177	22.047630	2.688308	9.355363	2.692042	4.072759	6.581256
E	7.167875	25.417007	27.214131	3.449610	11.314350	3.154841	4.604898	7.242146
HR	2.664507	25.881872	30.950736	4.027710	12.390232	3.372662	4.887931	7.647172

Tackling Balanced Dataset Option

Problem: Data Imbalance

- Significantly higher number of completed loans relative to defaulted loans

Solution 1: Rebalance Dataset to 50/50

- Fault: Detrimental & causes a negative average return
- Reason: More losses on investments as opposed to gains

LendingClub's Solution

- Continue analysis using unbalanced dataset
- Incorporate calibration curve into model building phase

Transforming & Saving Dataset

- Categorical variables used as is
- Numerical columns converted into continuous features
- Dataset is saved using pickle & exported as .CSV

Solution 1's Main Shortcoming

	grade	default	int_rate	return_PESS	return_OPT	return_INTa	return_INTb	return_INTc
A	19.973938	8.612483	9.474686	2.274137	4.541028	2.254222	3.486214	5.706550
AA	10.377824	4.014627	6.775790	2.019641	3.706281	2.023254	3.275717	5.534159
B	22.979313	12.956079	12.394934	2.395357	5.703969	2.466247	3.724494	5.997924
C	24.597005	18.515979	16.501650	2.495612	7.242789	2.583270	3.886307	6.246494
D	12.239538	23.292177	22.047630	2.688308	9.355363	2.692042	4.072759	6.581256
E	7.167875	25.417007	27.214131	3.449610	11.314350	3.154841	4.604898	7.242146
HR	2.664507	25.881872	30.950736	4.027710	12.390232	3.372662	4.887931	7.647172

Unbalanced Dataset Returns

	grade	default	int_rate	return_PESS	return_OPT	return_INTa	return_INTb	return_INTc
A	16.375878	34.944187	9.538783	-2.474031	-0.878019	-0.910899	0.147415	2.054459
AA	7.315120	18.946001	6.812487	-0.541292	0.901011	0.408420	1.568575	3.660394
B	21.652157	45.740079	12.460648	-3.750521	-1.679114	-2.000569	-0.990740	0.833388
C	26.888167	56.344973	16.524345	-4.971645	-2.390153	-3.167542	-2.191823	-0.425022
D	15.023438	63.123814	21.864717	-6.024556	-2.782839	-4.024851	-3.049293	-1.277420
E	9.258488	65.457922	26.980508	-6.464774	-2.762397	-3.933130	-2.920515	-1.078863
HR	3.486752	65.792889	30.860243	-7.820892	-3.933467	-4.032661	-2.986616	-1.081351

Balanced Dataset Returns

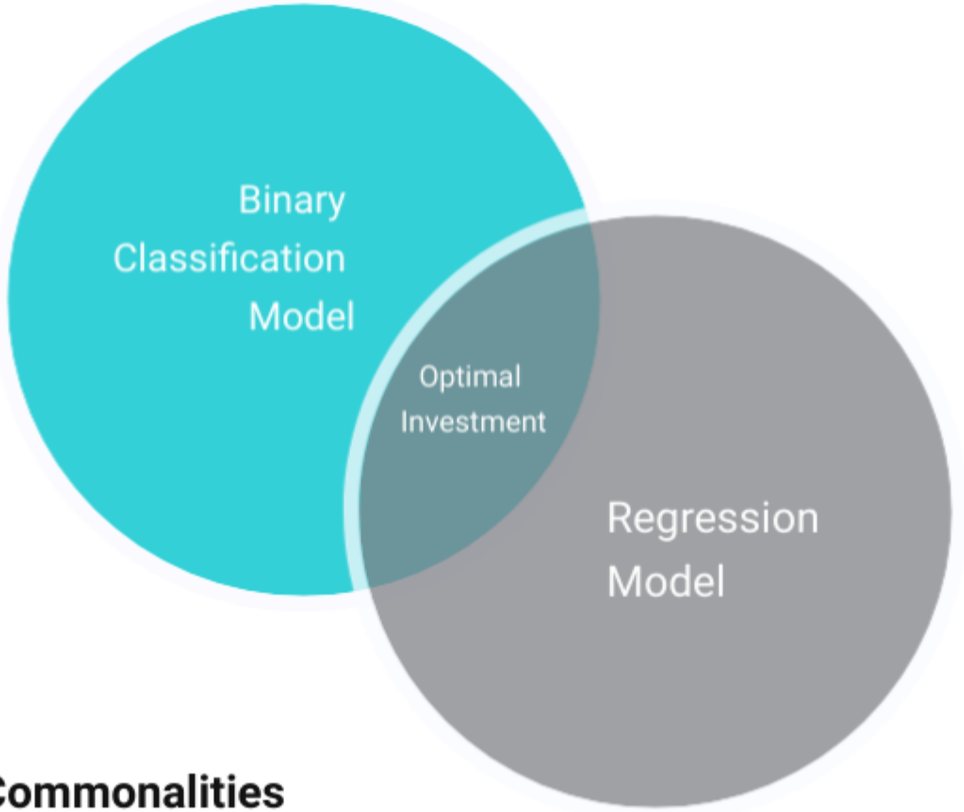
Predictive Model

Part 4

2 Phase Modelling

Binary Classification Model

- Predict loan default probability
- Use variety of industry standard algorithms
- Evaluate performance to find optimal model



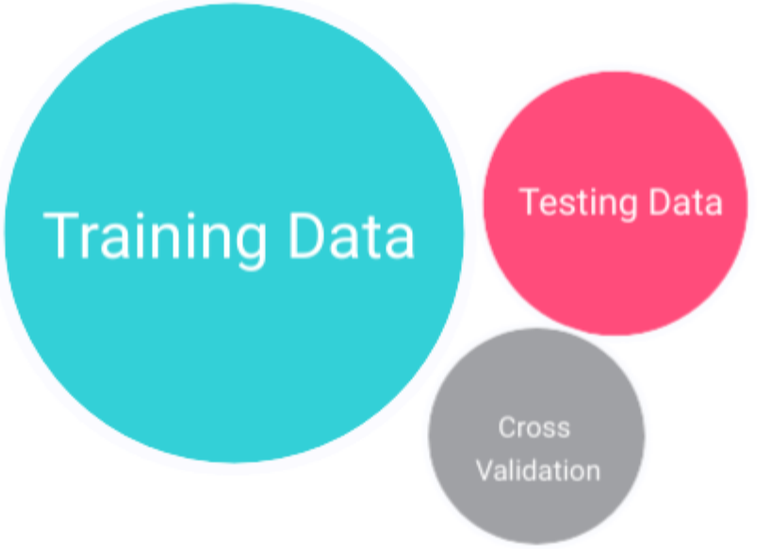
Regression Model

- Use variety of industry standard regressors
- Predict amount of return a loan may generate to investor

Commonalities

- Optimization for higher accuracy through CV's hyper-paramater tuning
- Models: Decision Tree, Random Forest, Logistic Regression, Naive Bayes, Multilayer Perceptron

Train / Test Split



Aim

- Fit a variety of models on various time periods
- Ensure each period uses the same training / test split
- Random assignment of each loan into the training & testing set
- Random Method Train Test Split:

Reasoning

- As per LendingClub case
- Easier to implement using standard libraries
- Allows thousands of train & test sets to be created using different random seeds

Dummifiying Variables & Model Function Definition

- All categorical variables
- Several functions (prepare_data, fit_classification, fit_regression, test_investments)
- Define few input parameters (model_name, cv_parameters) & search for optimal hyperparameters
- test_investments function enables data to be fit through a classification & regression model & returns using test dataset

Model Evaluation Metrics

Accuracy

- Not a strong determinant on its own, especially with an imbalanced dataset

Solution 1: Precision

- Correctly identifying a "good" loan

Solution 2: Recall

- Correctly identifying a "bad" loan

Trade-off between two: need to find the right balance depending on objective

F1-Score

- Weightage average of precision (p) and recall (r)
- Crucial measure in imbalanced datasets

ROC/AUC Curve

- ROC: plots true positive & false positive rate
- AUC: aggregate measure of performance across all possible classification thresholds

Calibration Curve

- Extent to which probabilities predicted by model correspond to frequency of event happening
- Ideal curve: 45-degree straight line

Grade & Interest Classification Model

2 Logistical Classification & Regression Models

- Using only grade / interest predictors
- Gave an AUC of 0.65
- Accuracy: 85% - industry standard

Results

- Same predictive power as when all predictors are used

Dropping all predictors derived by grade / interest results:

- Consistent with LendingClub where accuracy remains very high even when grade & interest predictors are dropped (roughly 85%)

Our Model

Target Variable

loan_status_description

Predictors

co_borrower application

term

prosper_fees_paid

service_fees_paid

amount_borrowed

Algorithms

Ridge Classifier

Naive Bayes

l2 Penalized Log. Reg.

Decision Tree

Random Forest

Bagged Tree

Multi-layer perceptron

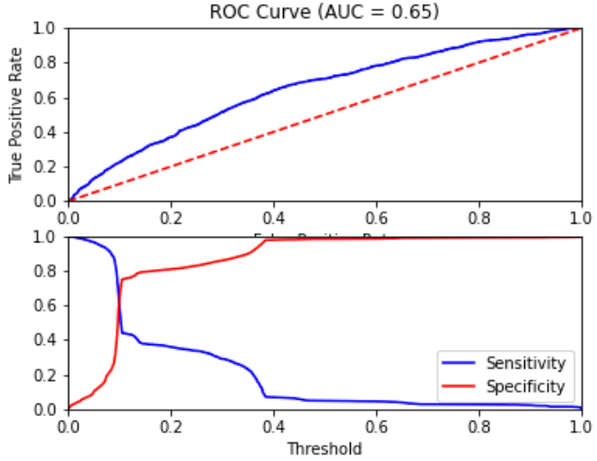
Our Additional Algorithm

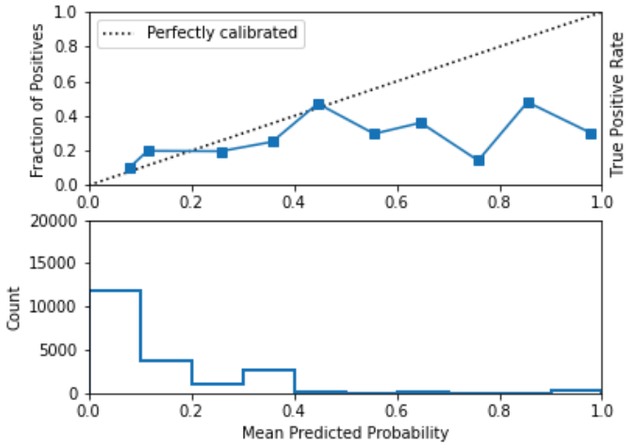
LightGBM*

Classification Results Comparison

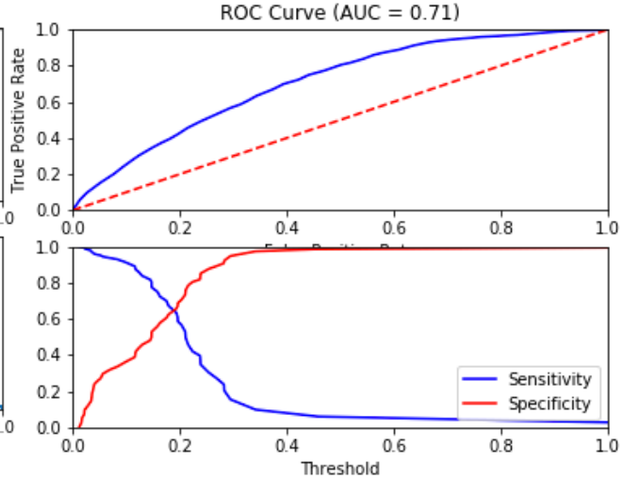
PROSPER

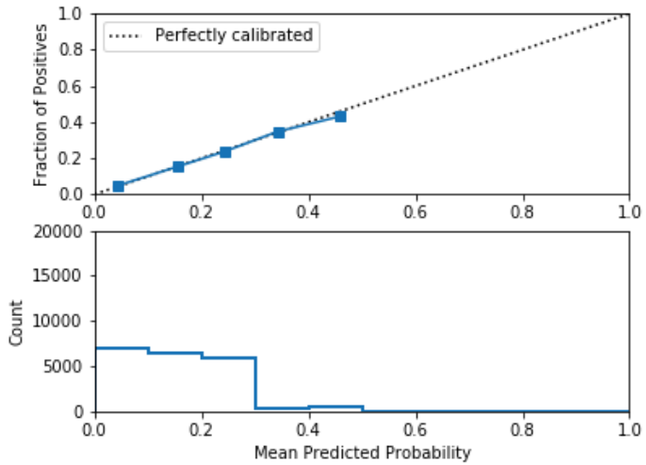
Naïve Bayes



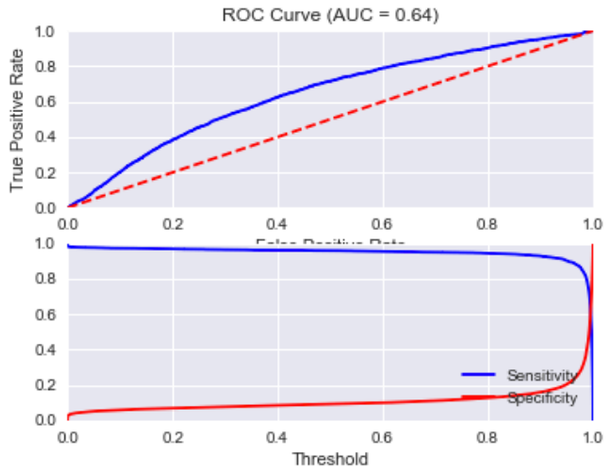


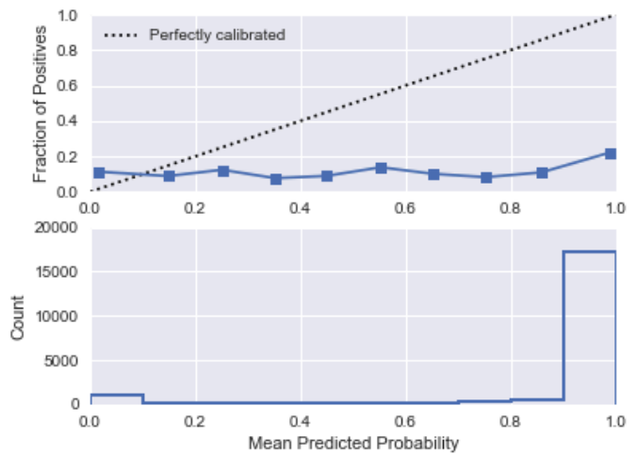
Decision Trees

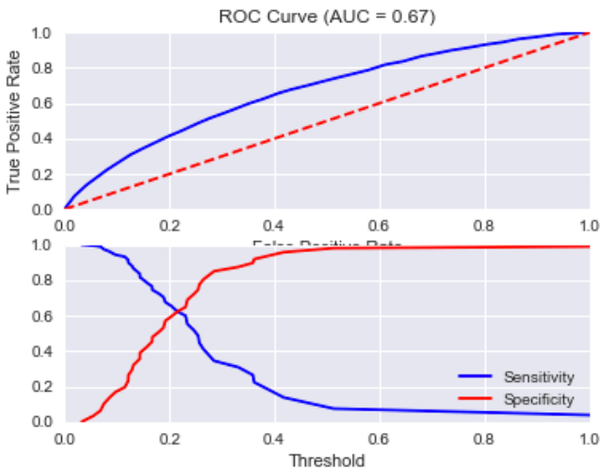


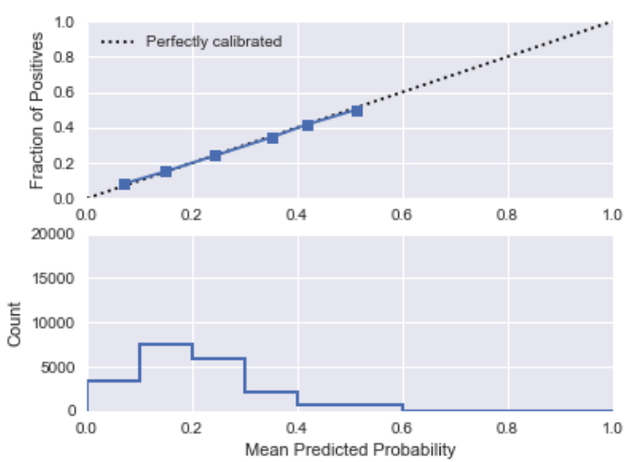


LendingClub





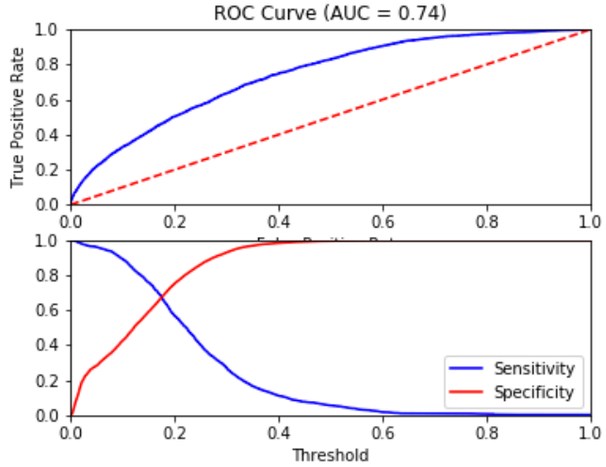


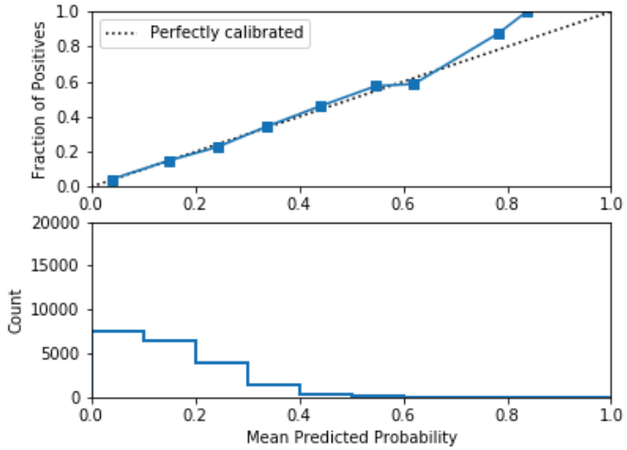


Classification Results Comparison

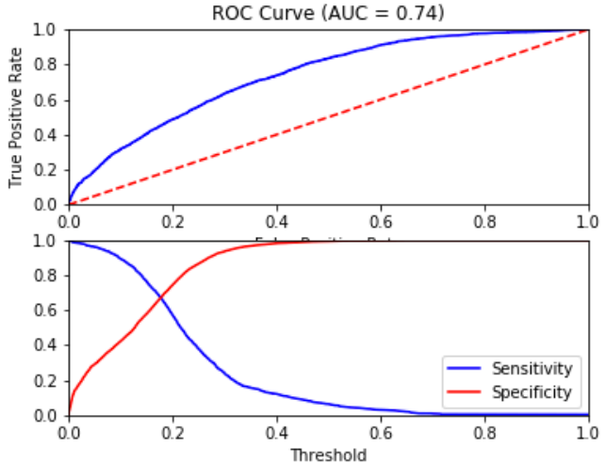
PROSPER

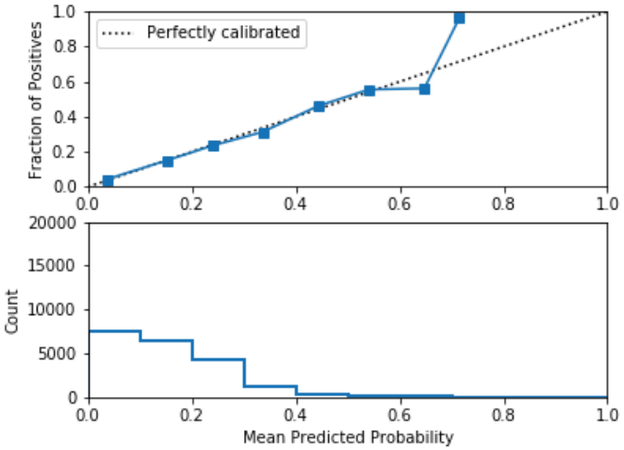
Random Forest



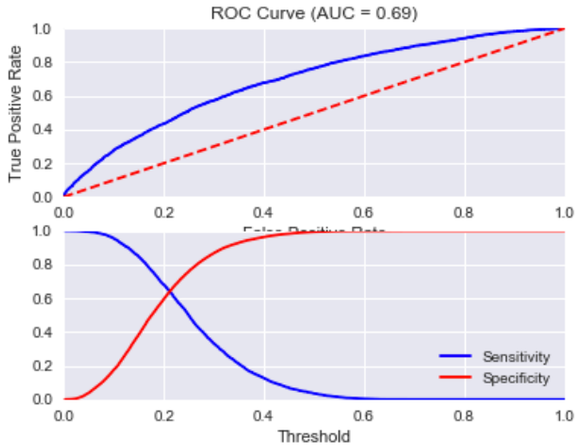


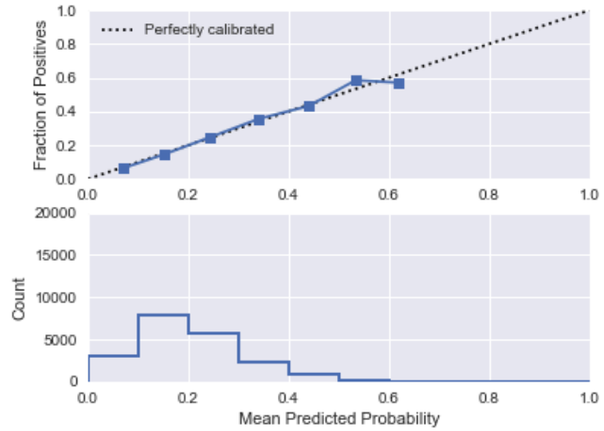
Bagging

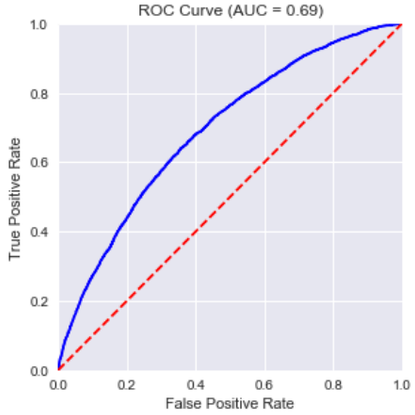


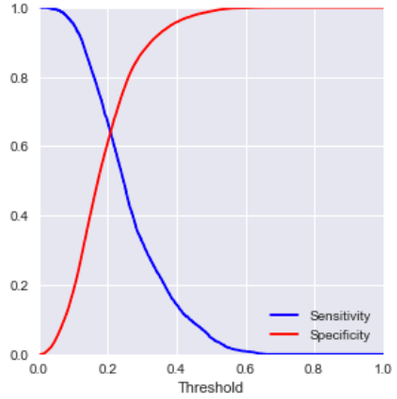


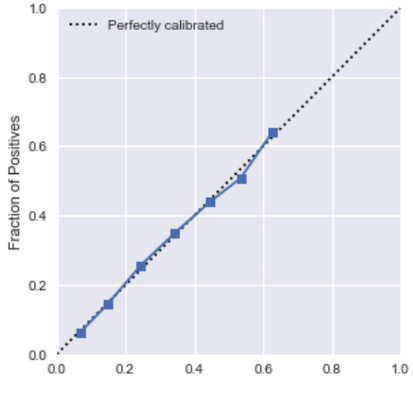
LendingClub











Classification Results

Classification Model	Accuracy	Precision	Recall	F1-Score	AUC	Calibration Level
Ridge Classifier	0.8489	0.8489	1	0.9183		
Naive Bayes	0.8489	0.8489	1	0.9183	0.65	Medium
l2 penalized logistic regression	0.8491	0.8493	0.9996	0.9184	0.64	Perfectly Calibrated
Decision tree	0.8489	0.8489	1	0.9183	0.71	Perfectly Calibrated
Random forest	0.8498	0.8629	0.9785	0.9171	0.74	High
Bagged trees	0.85195	0.8573	0.9905	0.9191	0.74	High
Multi-layer perceptron	0.85075	0.8566	0.99	0.9184	0.73	High
Light GBM	0.85185	0.8566	0.9914	0.9191	0.74	High

Regression Model Analysis

The objective of building the regression models was to predict/calculate the return for different return scenarios

The target variable are broken into the 4 types of return possible and have the regression model run on each of them separately: Pessimistic Approach (M1), Optimistic Approach (M2), ret_Inta (M3), ret_INTb (M3)

We did not use the ret_intC variable since our objective was to capture the occurrence of increase in the interest rather than the actual increase by value.

Regression Results

LendingClub

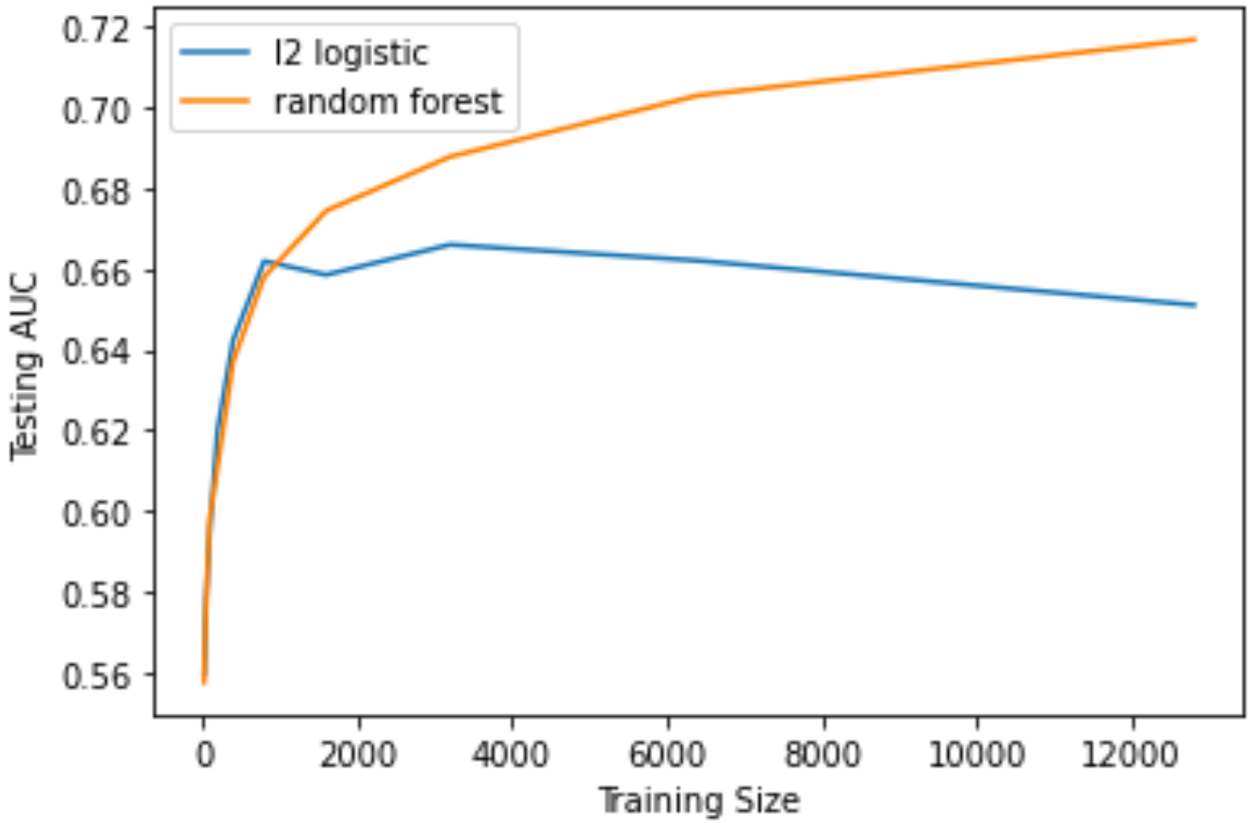
Model	R^2 scores for each return definition			
	M1	*M2*	*M3 (1.2%)*	*M3 (3%)*
ℓ_1 regressor	0.026	0.012	0.027	0.027
ℓ_2 regressor	0.026	0.013	0.027	0.027
Multilayer perceptron regressor	0.016	0.004	0.017	0.017
Random forest regressor	0.028	0.018	0.029	0.032

PROSPER

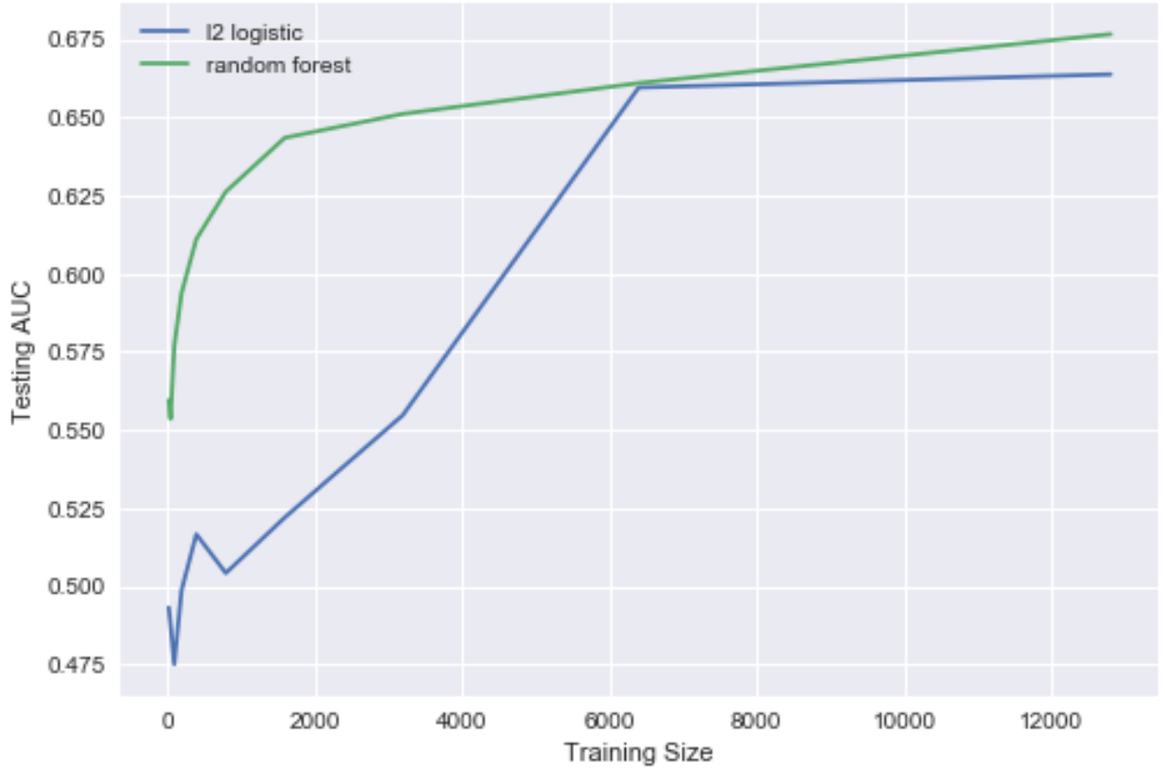
Model (Regressor)	R^2 scores for Each return definition			
	M1: Pessimistic	*M2: Optimistic*	*M3: (1.2%)*	*M3: (3%)*
Ridge	0.176	0.025	0.186	0.153
Ordinary Least Squares	0.176	0.025	0.186	0.153
Multi-Layer Perceptron	0.239	0.029	0.255	0.175
Random Forest	0.313	0.066	0.288	0.245

Learning Curve

PROSPER



LendingClub



Investment Strategies

Accuracy

Leveraging both the classification and regression models to predict whether a loan will get defaulted and the estimated return, an investment strategy can be formulated to maximize an investor's average return.

LendingClub

	Return calculation method			
	M1-PESS, %	*M2-OPT, %*	*M3 (1.2%)*	*M3 (3%)*
Rand	0.6	5.2	1.2	2.7
Def	1.9	5.3	1.4	3.0
Ret	2.6	5.6	1.6	3.2
DefRet	3.0	5.7	1.7	3.3
Best possible	12.0	27.1	10.1	11.7

4 Strategies

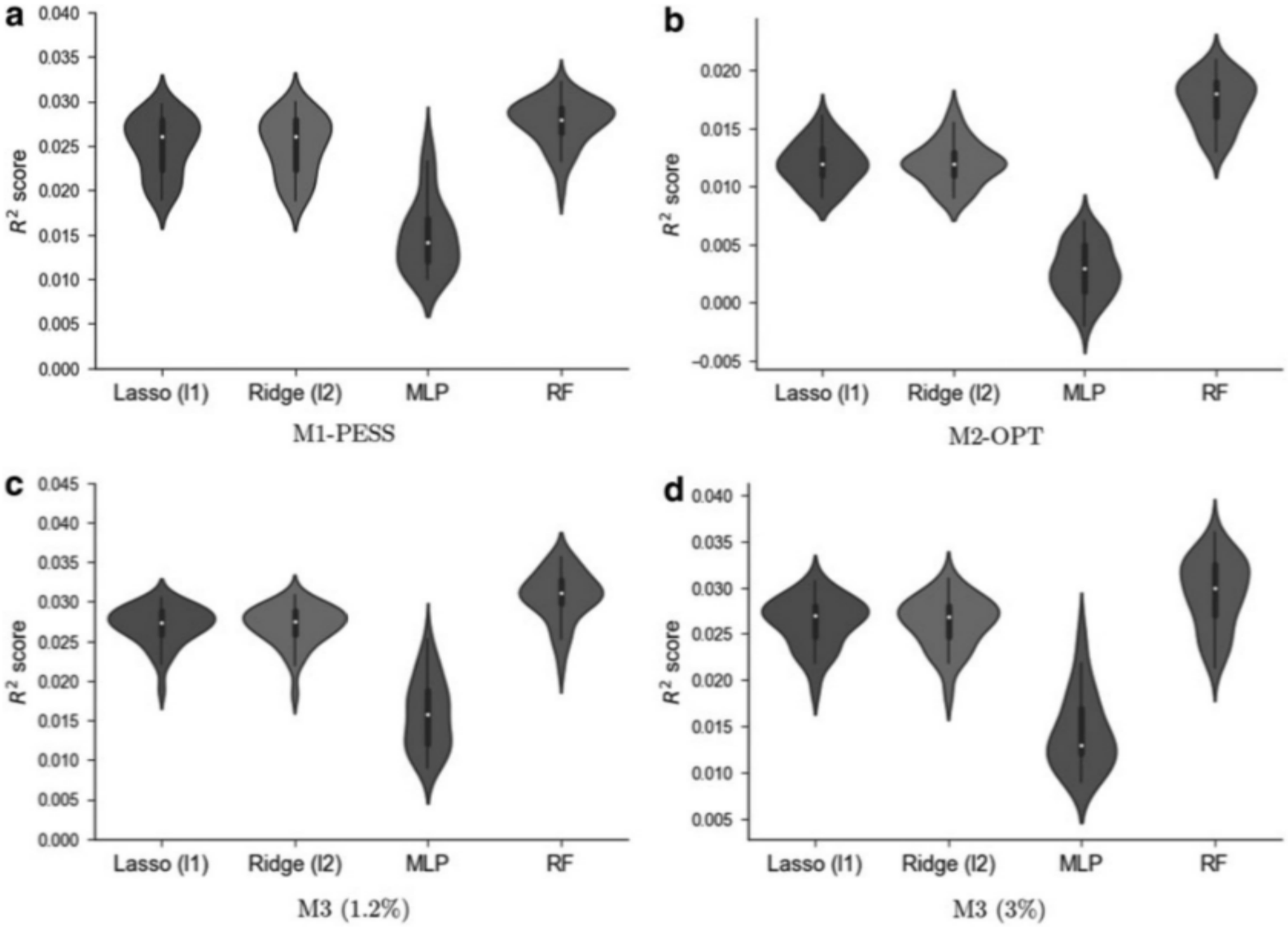
1. Random strategy (Rand)
2. Default based strategy (Def)
3. Simple return based strategy (Ret)
4. Default and return based strategy (DefRet)

PROSPER

Investment Strategy	Pessimistic Return (M1) %	Optmistic Return (M2)%	M3 (1.2%)	M3 (3%)
Rand	1.5	5.7	2.5	3.4
Def	7.3	7.5	2.7	3.2
Ret	1.1	7.3	2.6	3.5
DefRet	1.2	6.3	2.7	3.3

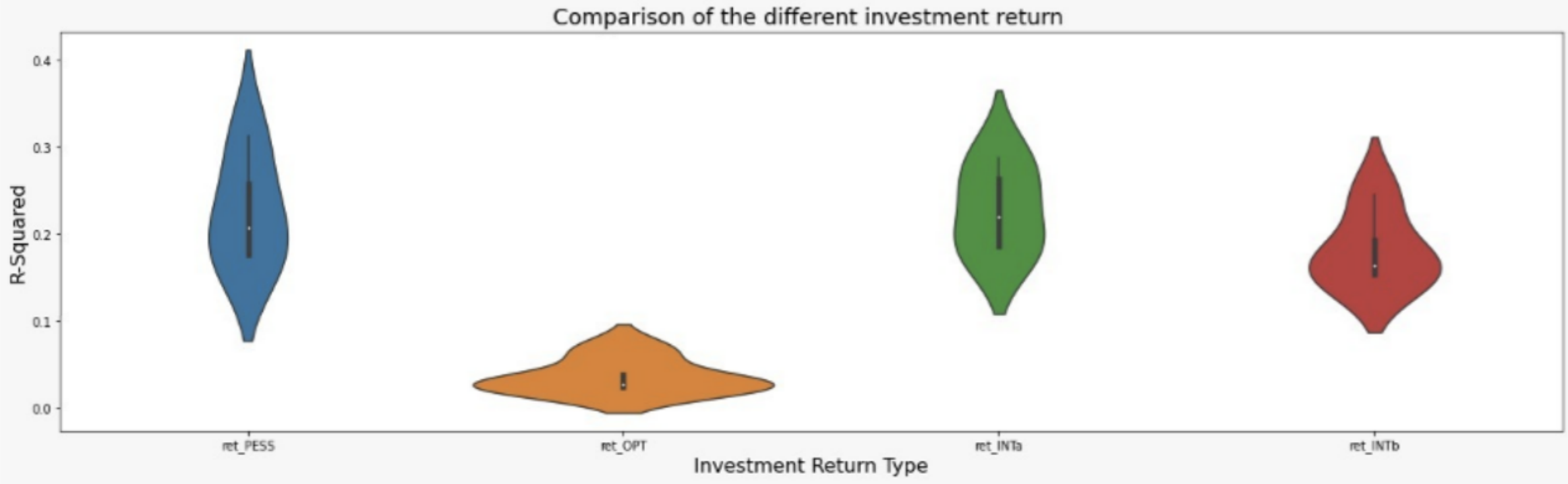
LendingClub

Comparison of Different Regression Models



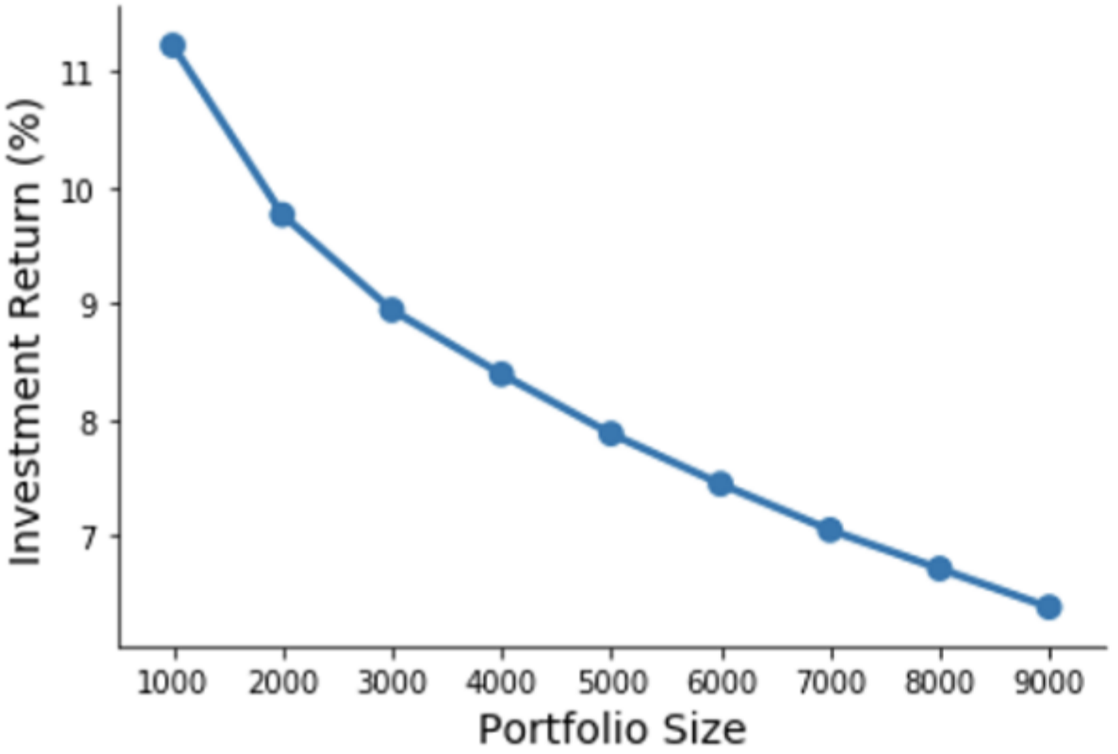
PROSPER

Comparison of Different Regression Models

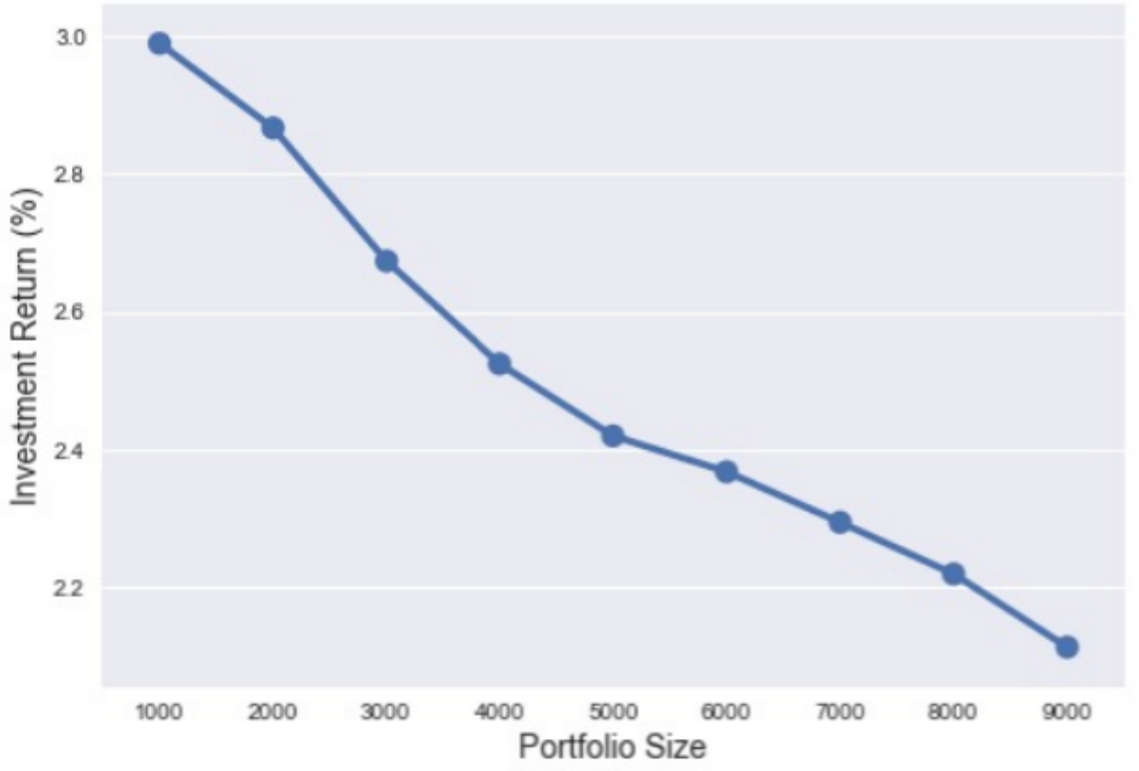


Sensitivity Plot

PROSPER



LendingClub



3 Optimization Models

1st Model

Maximizing expected profit

Binary variables

Number of loans constraint

LendingClub's optimal % of return: 2.59%

Prosper's optimal % of return: 5.8%

2nd Model

Budget constraint of potential investor

Budget limiting constraint added

Test different budgets

LendingClub's optimal % of return: 3.18%

Prosper's optimal % of return: 10.5%

3rd Model

Incorporating portfolio risk factor

Consider variance of returns

Clustering model trained with adjustable k parameter

Cluster based on Euclidean distance from each cluster

Sensitivity factor can be set by investor to account for investor's risk tolerance

LendingClub's optimal % of return: 3.21%

Prosper's optimal % of return: 10.3%

Conclusion

Prosper is a newer P2P lending platform

- Lacks the stability that LendingClub's data has cultivated over time
- Uncertainty regarding our model's future applicability
- Prosper's business model differs as a result of having higher credit scores requirements & fees
- LendingClub possesses a more defined investment approach

Prosper ≠ LendingClub

- Both lenders have different methodologies in setting their interest rates
- There is no "one best lender" as it is highly dependent on the investor's needs