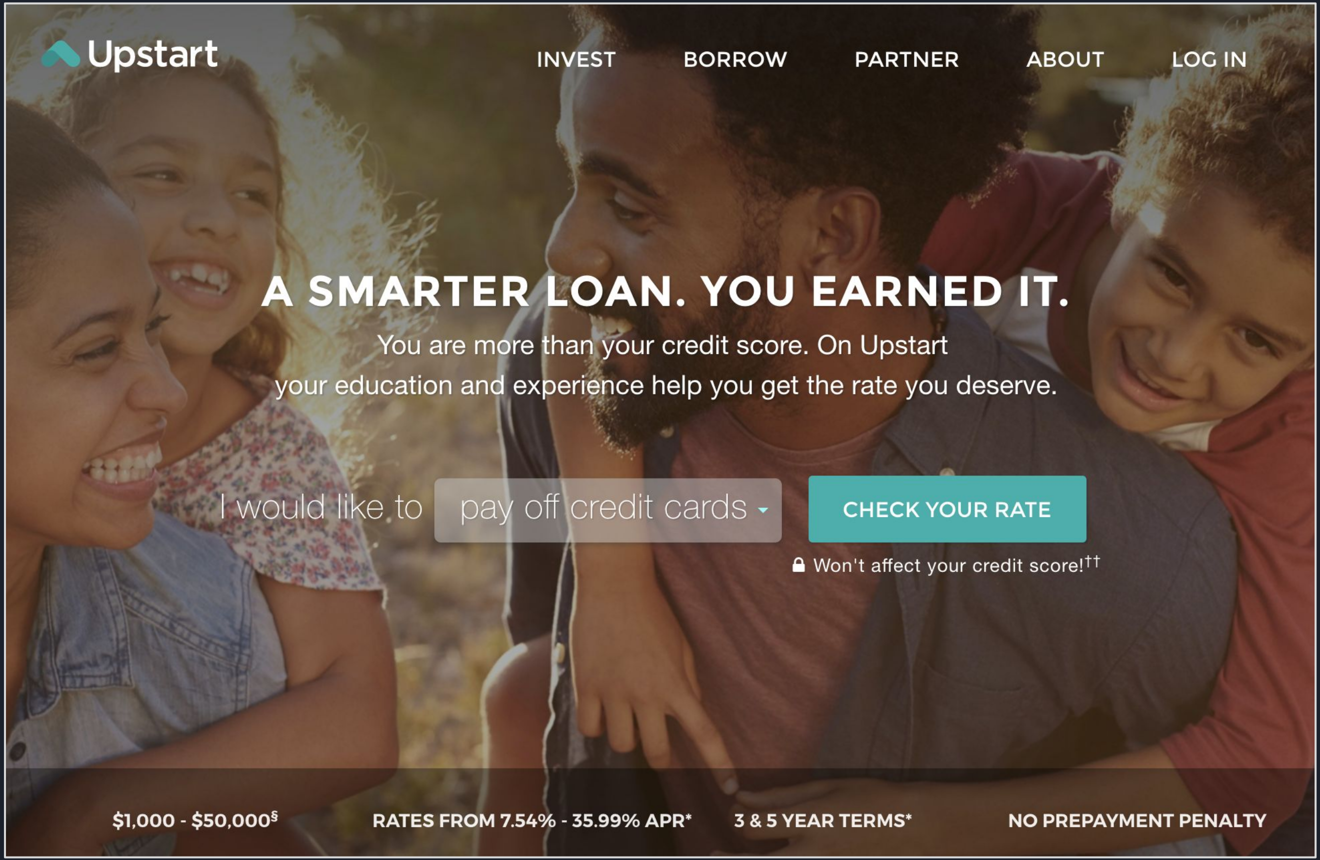
Predicting Customer Behavior: Who will take out an Upstart loan?

Farouk Rice

Upstart spends thousands of dollars a month on advertising to loan applicants.

What if we could predict which applicants were actually likely to take out a loan?

The Product


Upstart
INVEST
BORROW
PARTNER
ABOUT
LOG IN
A SMARTER LOAN. YOU EARNED IT.
You are more than your credit score. On Upstart
your education and experience help you get the rate you deserve.
I would like to
pay off credit cards
CHECK YOUR RATE
Won't affect your credit score!††
$1,000 - $50,000§
RATES FROM 7.54% - 35.99% APR*
3 & 5 YEAR TERMS*
NO PREPAYMENT PENALTY

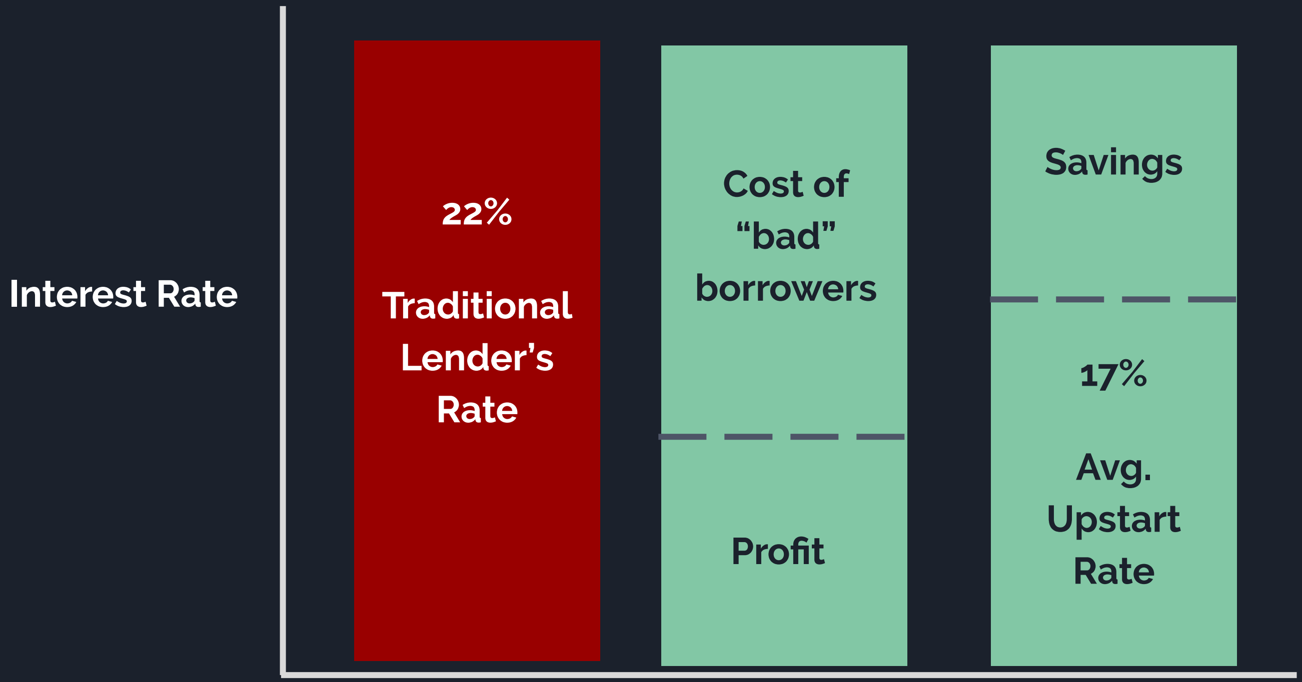
The Product

Income

Credit Score		$20000 - $40000	$40000 - $80000	$80000+
	680 - 720	35% APR	30% APR	25% APR
	720 - 760	30% APR	25% APR	20% APR
	760+	20% APR	15% APR	10% APR

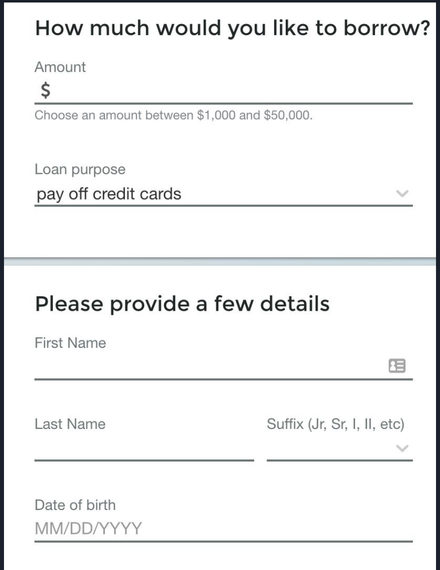
APR = Annual Percentage Rate, or the “price” of the money borrowed

The Product


Interest Rate
22%
Traditional Lender's Rate
Cost of "bad" borrowers
Profit
Savings
17%
Avg. Upstart Rate

The Customer Funnel

Application



How much would you like to borrow?

Amount

$

Choose an amount between $1,000 and $50,000.

Loan purpose

pay off credit cards

Please provide a few details

First Name

Last Name

Suffix (Jr, Sr, I, II, etc)

Date of birth

MM/DD/YYYY

~27%*

Verification

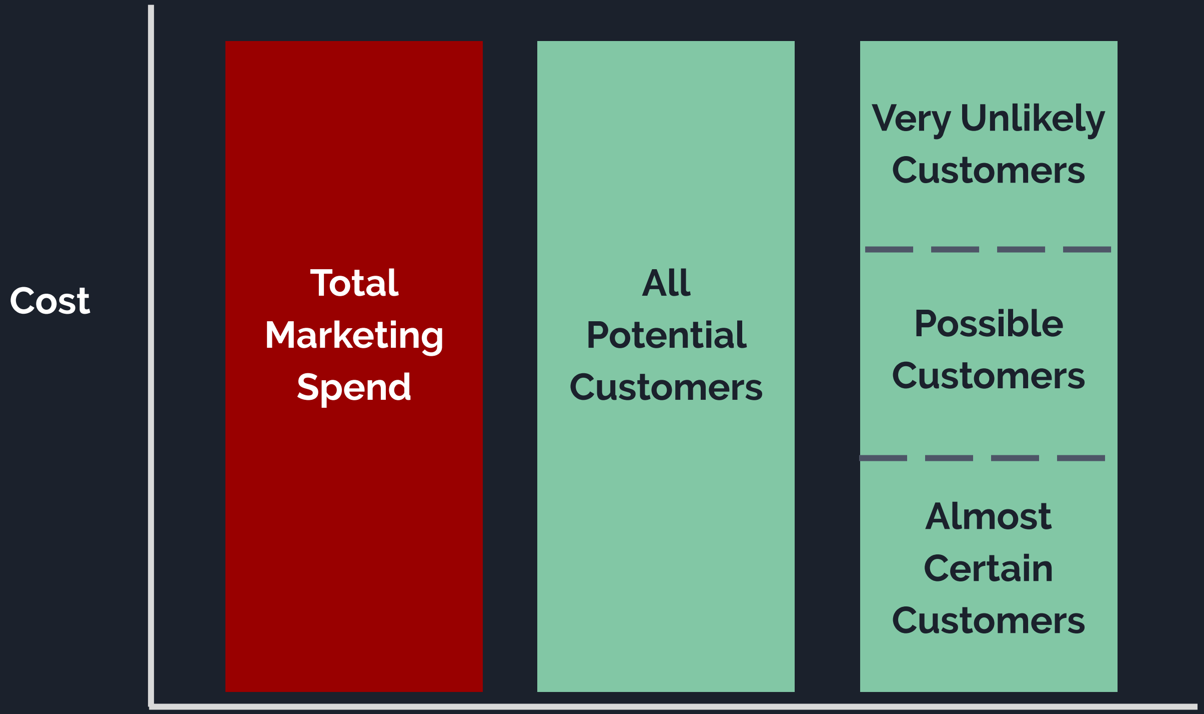
~71%

Acceptance

19% overall

* 27% of eligible applicants

The Potential



The Potential

Remarketing costs:
Approx. $200,000 /mo

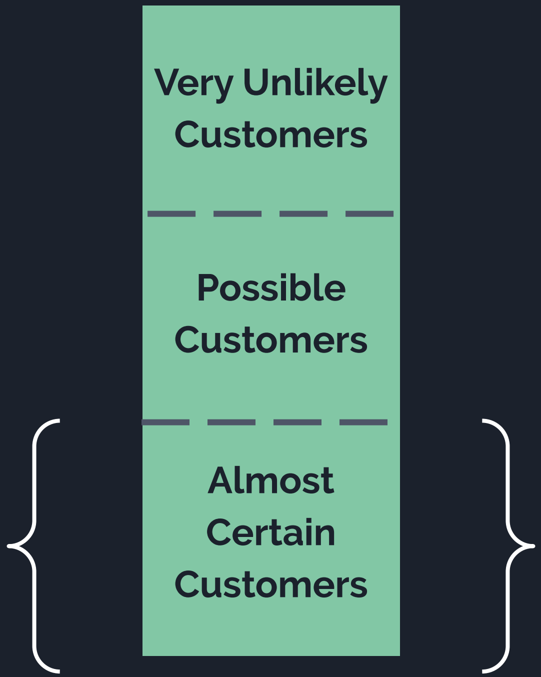
Very Unlikely Customers

Possible Customers

Almost Certain Customers

Revenue:
Approx. $3MM /mo

The Potential



Remarketing costs:
~19%, or $38K/mo

Revenue:
Approx $3MM /mo

The Potential

Remarketing costs:
No add’l cost

Very Unlikely Customers

Possible Customers

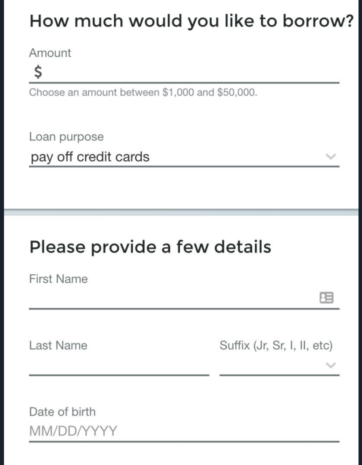
Almost Certain Customers

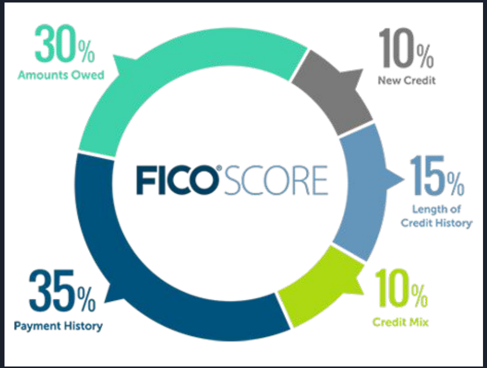
Revenue:
Unknown increase

5% increase = $150K /mo

10% increase = $300K /mo

The Data



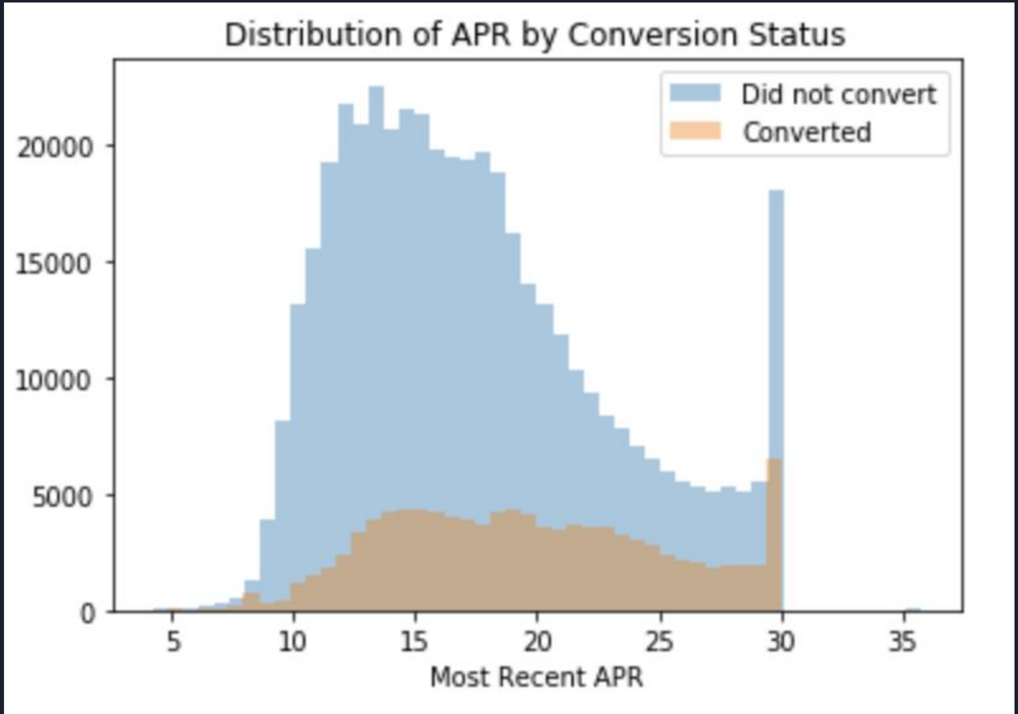


Upstart

Outcome variable:

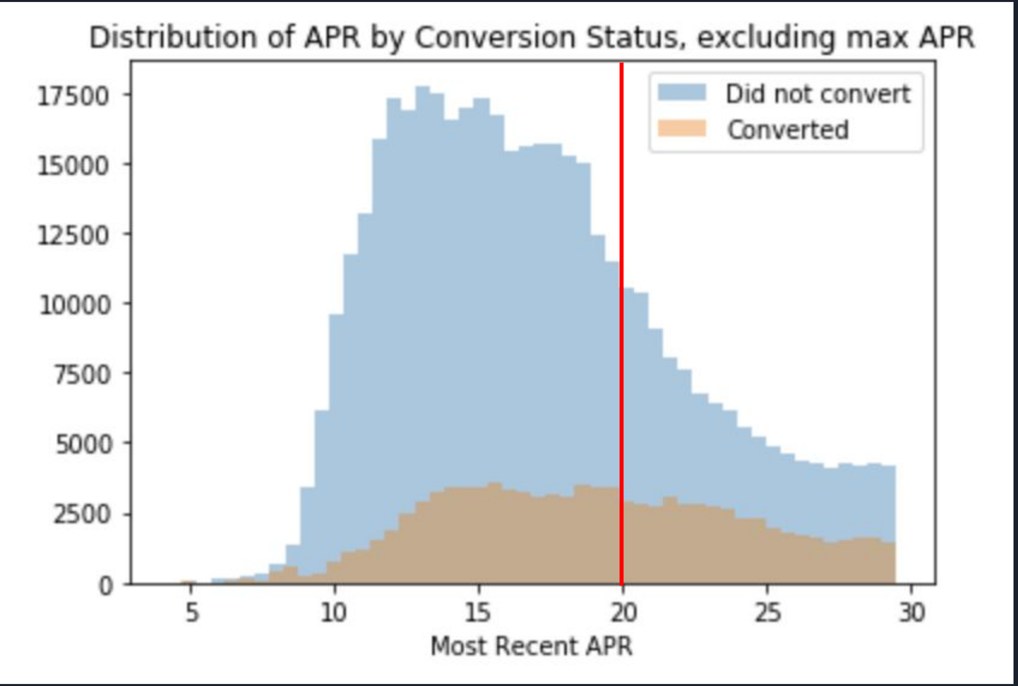
- Did the user take out a loan? (conversion)
  - Baseline: Among qualified applicants in the training set, 76.6% do not take out a loan.
  - Our model needs to be better than 76.6%.

Feature Exploration



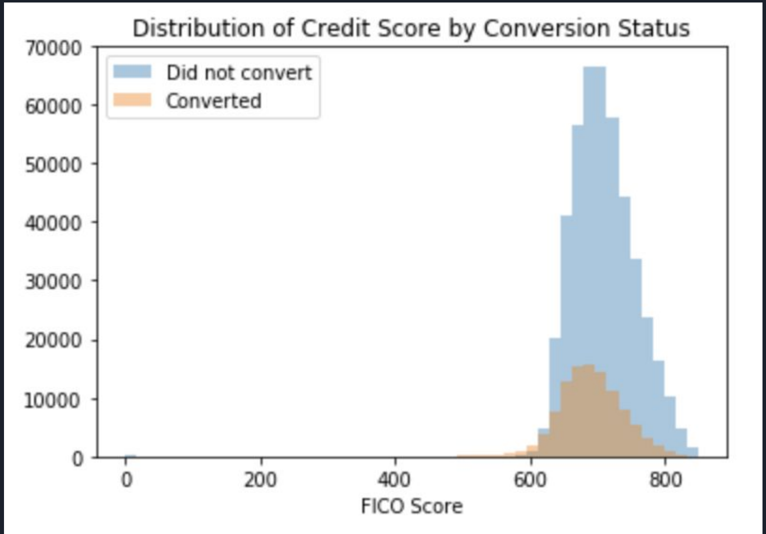
- Maximum APR of 30%
- Higher proportion of high APR users convert

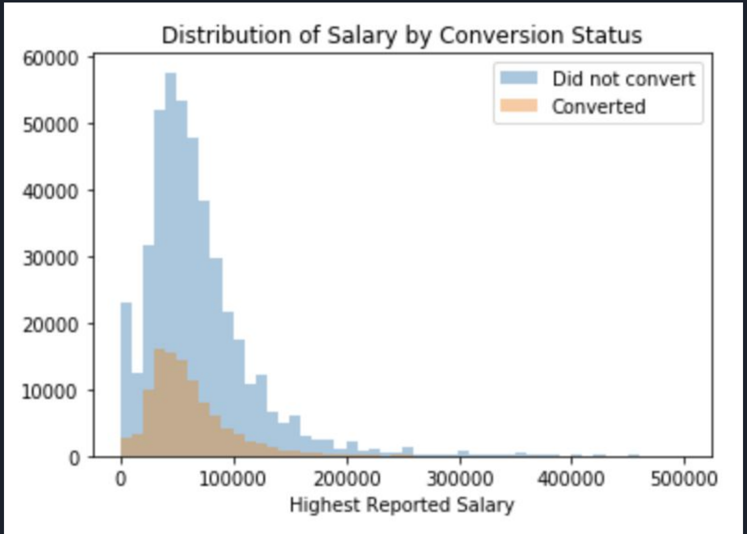
Feature Exploration



	Conversion Rate
Applicants above 20% APR	26.7%
Applicants under 20% APR	15.3%

Feature Exploration

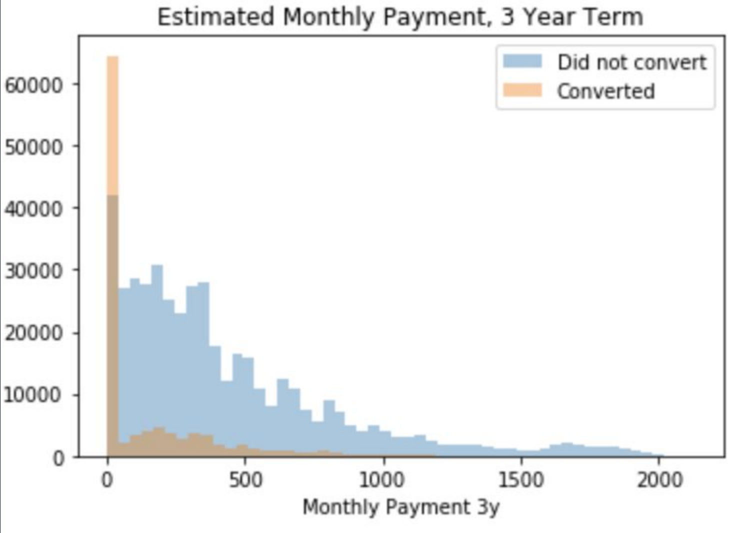


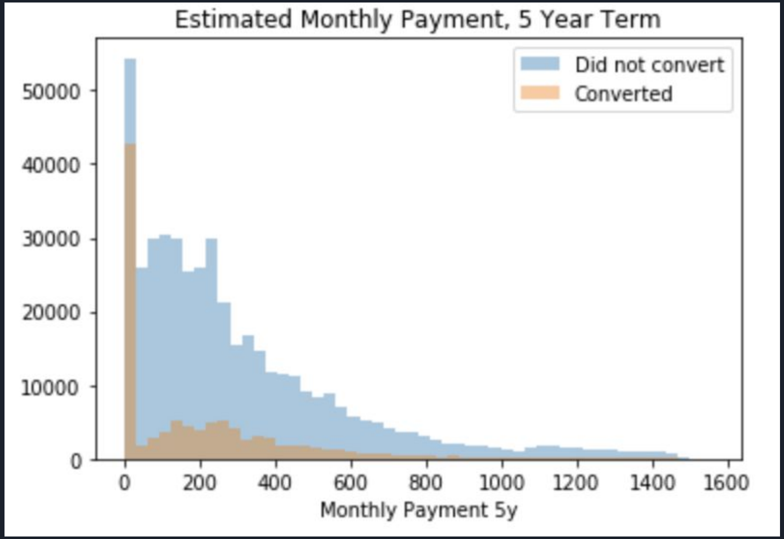


	Skewness
Did not convert	-1.74
Converted	-0.46

	Skewness
Did not convert	2.47
Converted	2.52

Feature Exploration





	Number of applicants	Conversion rate
Eligible for 3 year only	72768	71.7%
Eligible for 5 year only	88294	57.9%

Model Selection

Data Shape

- 958996 observations, 22 imported features

Logistic Regression

- Simple and easy to interpret
- Probably not robust enough to interpret complex features

Multi-Layer Perceptron

- Good with high number of observations
- Can find complex relationships between features

Random Forest

- Good with high number of observations
- Can find complex relationships between features
- Proven success in finance industry with credit scores and underwriting

About The Model

Hyperparameter-tuned Random Forest

- Roughly 1600 trees with a maximum depth of 100
- Better performance than linear regression and Multi-Layer Perceptron
- Features
    - Debt-to-income percentage
    - Loan amount requested (or highest amount approved for, whichever is higher)
    - Monthly Payment for 3 and 5 year terms
    - APR
    - Number of other loans
    - Credit Score
    - Reported Salary
    - Age
    - Likelihood of Fraud (Low, High)
- Iterations
    - One-hot encoded categorical variables
    - Time features
    - No categorical variables
    - Some categorical features
        - Hyperparameter tuning

About The Model: Predictive Power

Test-Train Split

- Before any exploration, 20% of the data reserved as test data
- Of remaining 80%, 20% was used as validation data and remainder used for training

	Validation Data Accuracy
Baseline	.81*
Logistic Regression	.852
Multi-Layer Perceptron	.811
Random Forest (final version)	.972
Random Forest (test data)	.971

*based on .19 overall conversion rate

About The Model: Predictive Power

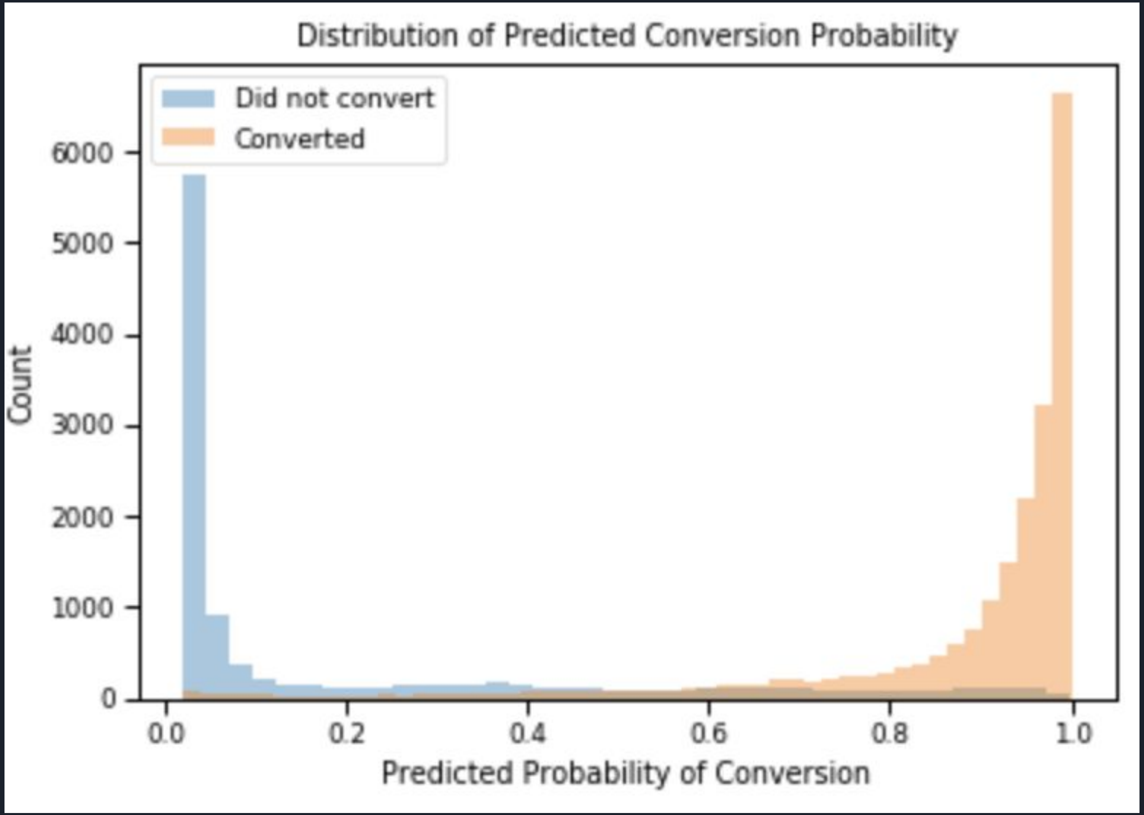
	Predicted non-conversion	Predicted conversion	Total
Actual non-conversion	109802	2390	112192
Actual conversion	1603	24980	26583
Total	111405	27370	138775

19% actual conversion rate overall, 19.7% predicted conversion rate

2.1% false positive rate (predicted conversion of non-converting customers)

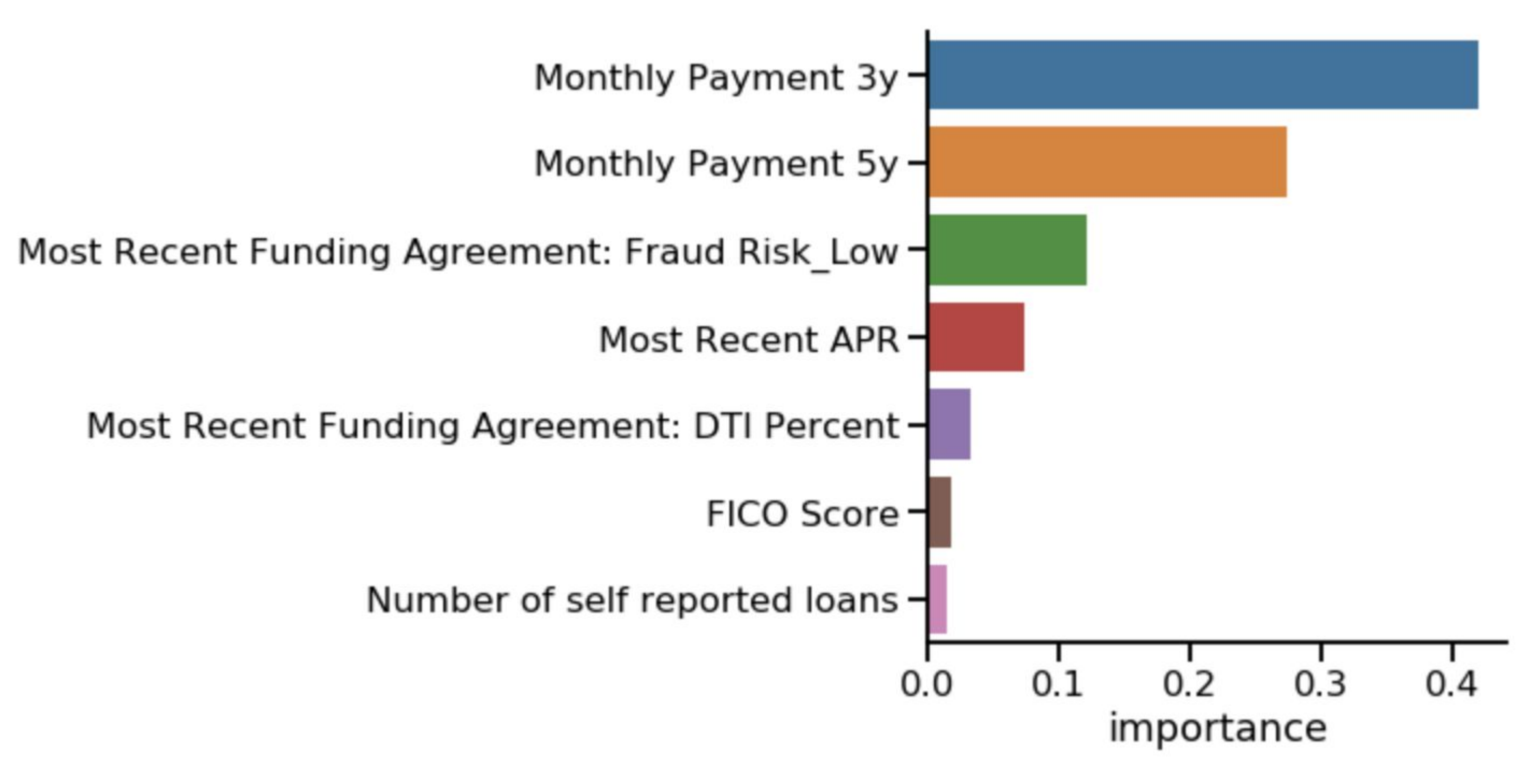
6% false negative rate (failed to predict conversion of converting customers)

About The Model: Predictive Power



- Removed 0% conversion probability
- Polarized assignment of probability
  - Relatively few in the middle

About The Model: Feature Importance

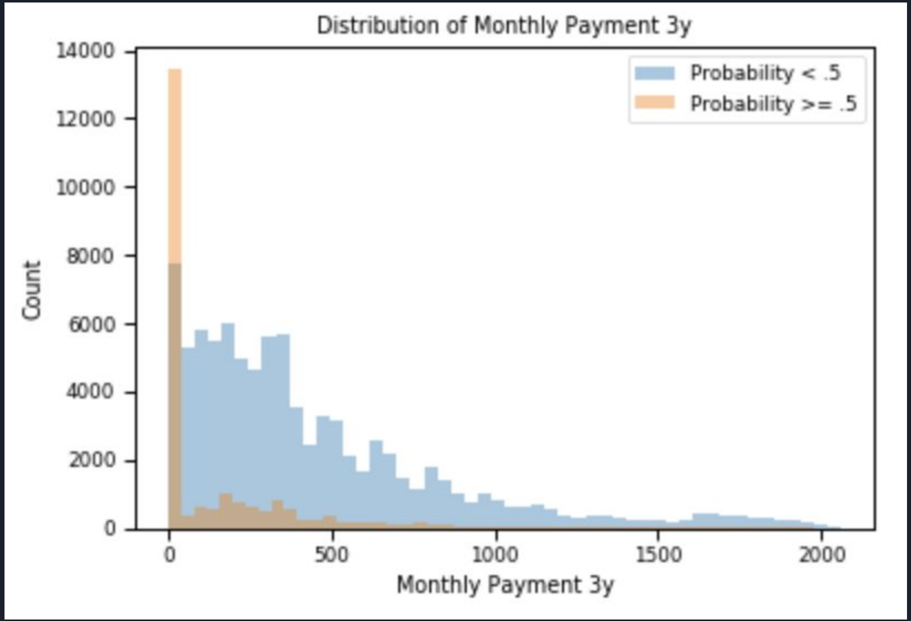


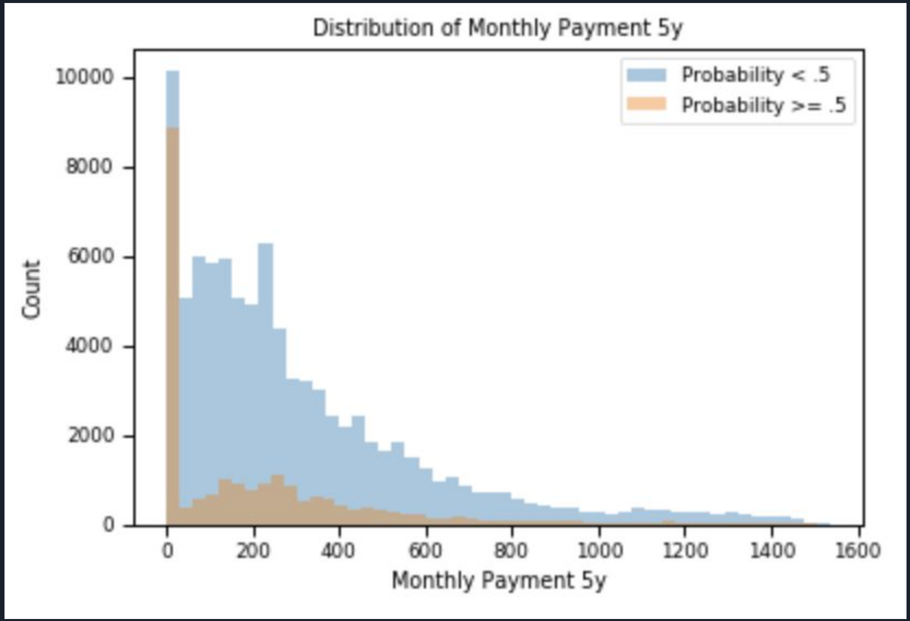
About The Model: Feature Importance

Logistic Regression largest coefficients:

Feature	Coefficient
Monthly Payment 5y	-.0084
Monthly Payment 3y	-.0081
FICO Score	.0015

About The Model: Feature Importance





- Highest probabilities given to those who qualify for exactly one loan term

What next?

Did not convert, but had 33% or higher probability:

Additional ~$60K in revenue

What next?

Very Unlikely Customers

Possible Customers

Almost Certain Customers

5% or lower probability:

Could save ~$126K in costs

95% or higher probability:

Could save ~$16K in costs

What next?

- Utilize Monthly Payment information to improve conversion rates
    - Less creditworthy borrowers generally more likely to convert
    - Further improve underwriting on these users or target these segments
    - Offer longer loan terms (lower Monthly Payments) to more people

Thanks!