Predicting Loan Defaults With Machine Learning

NYC Data Science Academy, April 2020 Cohort

Philippe Heitzmann

Disclaimer

The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the opinions, official policies or positions of any entities with which author has been, is now or will be affiliated.

Project Outline

1. **Why predict loan defaults?**
2. **Lending Club Background**
3. **Research Goals**
4. **Exploratory Data Analysis**
5. **Data Preprocessing**
6. **Feature Engineering & Selection**
7. **Classification Modeling with Machine Learning & Deep Learning**
8. **Results & Conclusion**

Why Predict Loan Defaults?

Key Uses

- Capital loss prevention & profit maximization for **private companies** extending credit to loan applicants
- Loss reserves and capital ratios forecasting for **financial regulators**
- Financial discrimination safeguards through interpretable models for **financial regulators**

Companies using Machine Learning for Loan Default Prediction

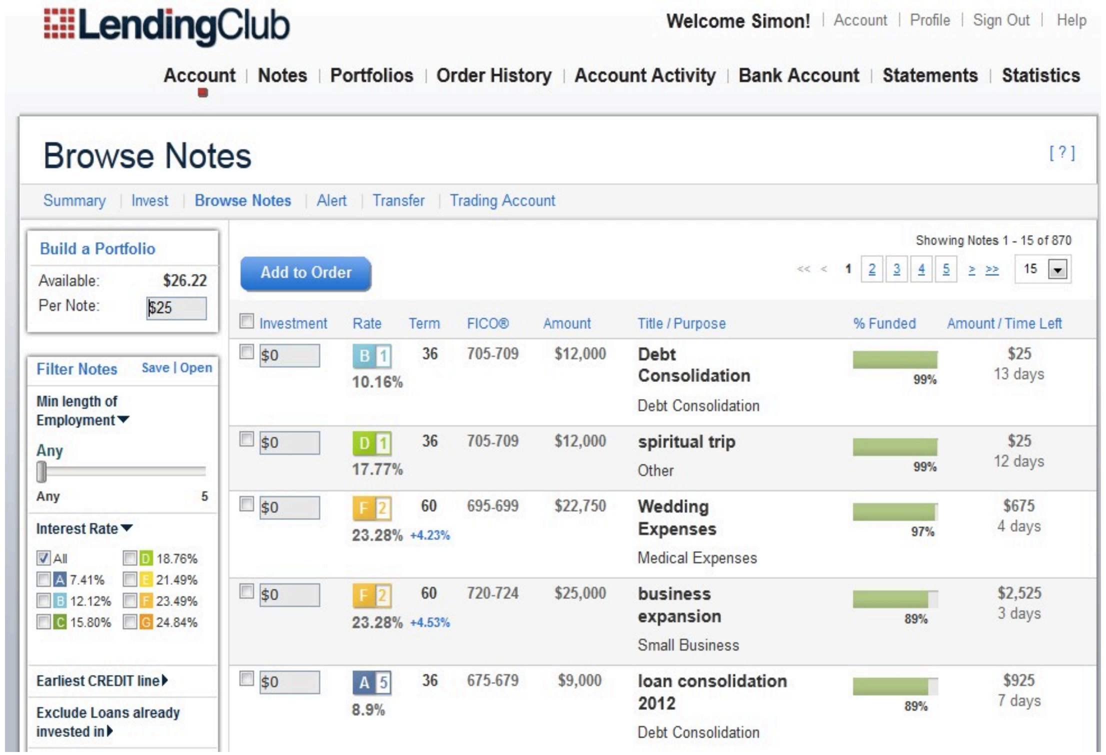
VISA · MasterCard · CHASE · ALLIANT · citi · AMERICAN EXPRESS · Capital One · DISCOVER · Bank of America

What is LendingClub ?

Background

- **Peer to peer (P2P)** financing platform
- Launched in 2006
- Allows investors to fund loans based on applicant's:
  - Credit history
  - FICO score
  - Employment status
  - Length of employment
  - Loan purpose
  - Loan grade
  - Other self-reported information

Sample Platform View



LendingClub Dataset Size and Completeness Allows for Machine Learning modeling

Dataset Overview

Description	Size
Unique Observations	2.3M
Number of Features	151
Feature Types	Numerical, qualitative, datetime
Dates spanned	2007 - 2018
Loan portfolio value	$34.01B
Missing values	731K (31.8%)

Key Considerations:

- **Class imbalance:** Non-defaulted loans outnumber defaulted loans **6.7 : 1**
- **Significant variable multicollinearity:** Certain variables exhibit high multicollinearity which may introduce modeling issues for linear models
- **Outliers:** Certain outliers may distort model coefficients and predictions and cause model to overfit

NYC DATA SCIENCE ACADEMY

Research Goals

1. **Produce machine learning and deep learning models trained on 2007-2017 data to accurately predict loan defaults in the 2018 loan pool**
2. **Optimize for the best investment opportunity set for an investor looking to maximize his or her returns on the 2018 loan set**
3. **Construct real-time machine learning prediction tools to allow investors to leverage classification models for portfolio construction**

Tools Used:

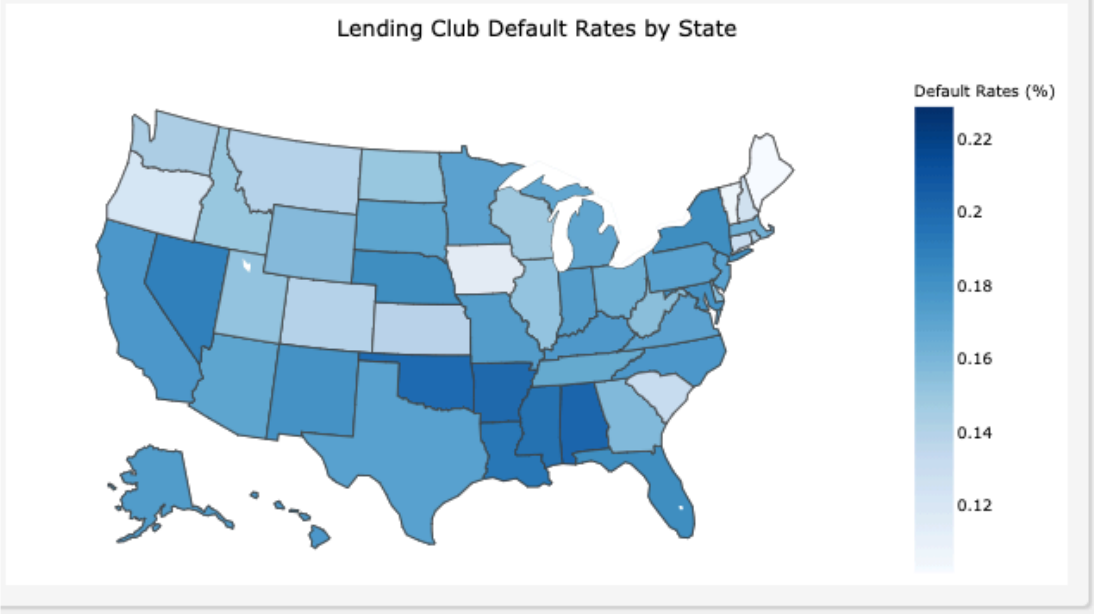
pandas, plotly | Dash, Scikit learn, Keras, Flask, NumPy, seaborn, CatBoost, TensorFlow, aws

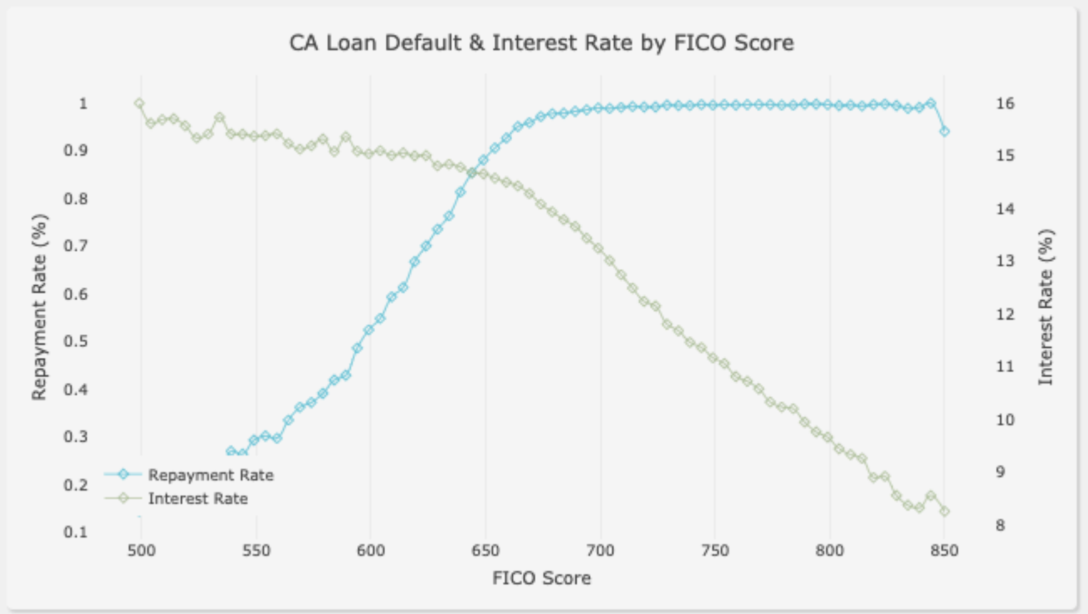
NYC DATA SCIENCE ACADEMY

Exploratory Data Analysis

Encourage everyone to follow along these Exploratory Data Analysis graphs at the following AWS instance link:

http://ec2-3-15-44-208.us-east-2.compute.amazonaws.com/





Data Preprocessing

Dropping variables that:

- **Contain >50% missing data**
  - **42 variables** dropped
- **Introduce data leakage**
  - **4 variables** dropped
- **Have only a single value**
  - **4 variables** dropped
- **Introduce random noise**
  - **variable** dropped
- **Display multicollinearity**
  - **variable** dropped

Imputing data using:

- **Mean** for numerical variables where most likely imputed value would be the average (annual income)
- **Zero** for numerical variables where missing value should indicate absence of a feature (employment length)
- **Highest-frequency class** for qualitative variables

Normalizing Features:

- Scaling training data to **[0,1] domain** to ensure better convergence for neural net
- Ensuring y target values are in the **[0,1] set**

Feature Engineering & Feature Selection

Goal: Provide investors with highest-Internal Rate of Return (IRR) for a loan subset

- Extracting datetime values from date variables passed in string format
- Writing function to output cash flow schedules for each loan
- Creating function to calculate net present value and IRR of a loan
- Create variable to capture misclassification costs to tune model classification cutoff threshold

Machine Learning & Deep Learning Modeling

Models Tested

Model Name	Model Type
Logistic Regression	Linear
Linear Discriminant Analysis	Linear
Quadratic Discriminant Analysis	Linear
Multinomial Naïve Bayes	Linear
Gaussian Naïve Bayes	Linear
Random Forest Classifier	Tree-ensembling
Gradient Boosting Classifier	Tree-ensembling
Catboost Classifier	Tree-ensembling
Neural Net	Deep Learning

Measuring Model Performance

ROC_AUC Score takes into account:

- Impact of false negatives to measure impact of lost interest income
- Impact of false positives to measure impact of loan principal loss

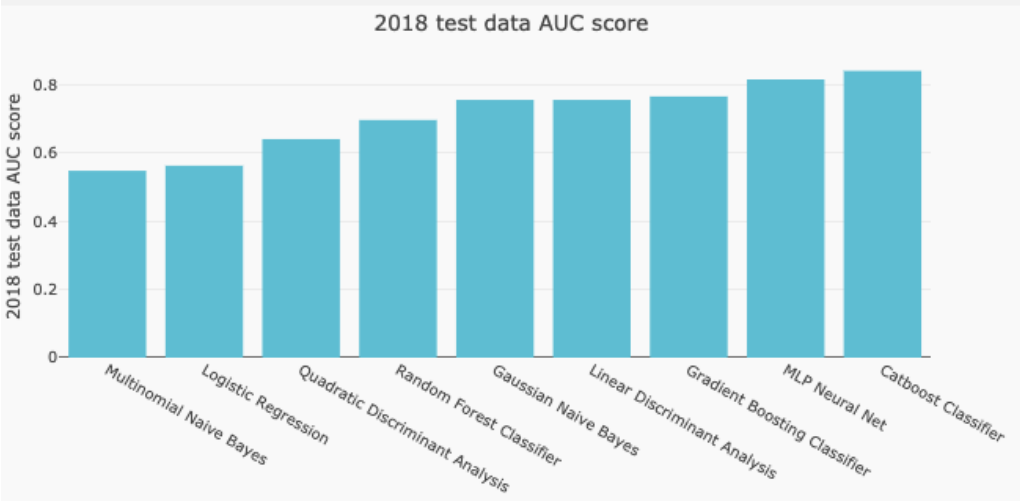
NYC DATA SCIENCE ACADEMY

Machine Learning & Deep Learning Results

Model Classification Performance

Model Name	2018 AUC Score
Catboost Classifier	0.841
Neural Net	0.816
Gradient Boosting Classifier	0.766
Linear Discriminant Analysis	0.757
Gaussian Naïve Bayes	0.756
Random Forests Classifier	0.697
Quadratic Discriminant Analysis	0.641
Logistic Regression	0.563
Multinomial Naïve Bayes	0.548

Side-by-side Model Comparisons

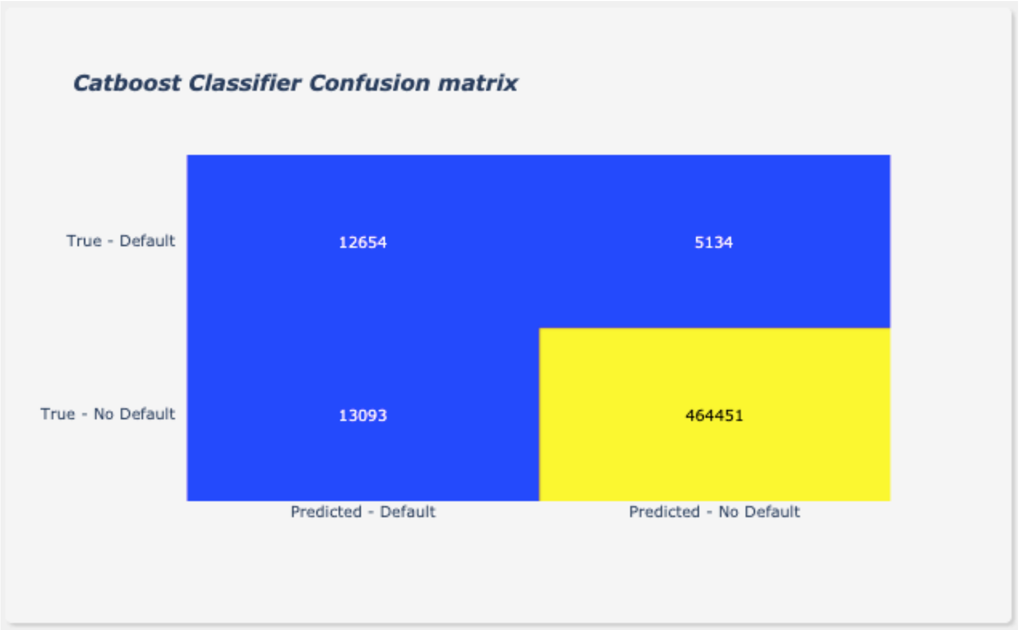


Takeaways:

- Tree-ensembling methods largely outperform linear models
- Random Forests surprising low-performer

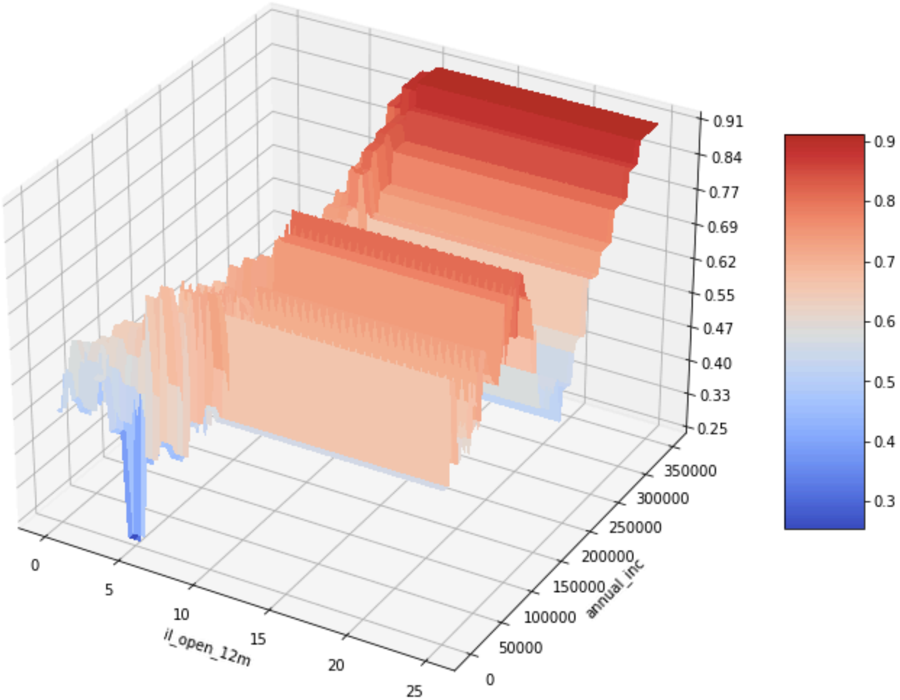
Visualizing Machine Learning & Deep Learning Results

CatBoost produces best results!



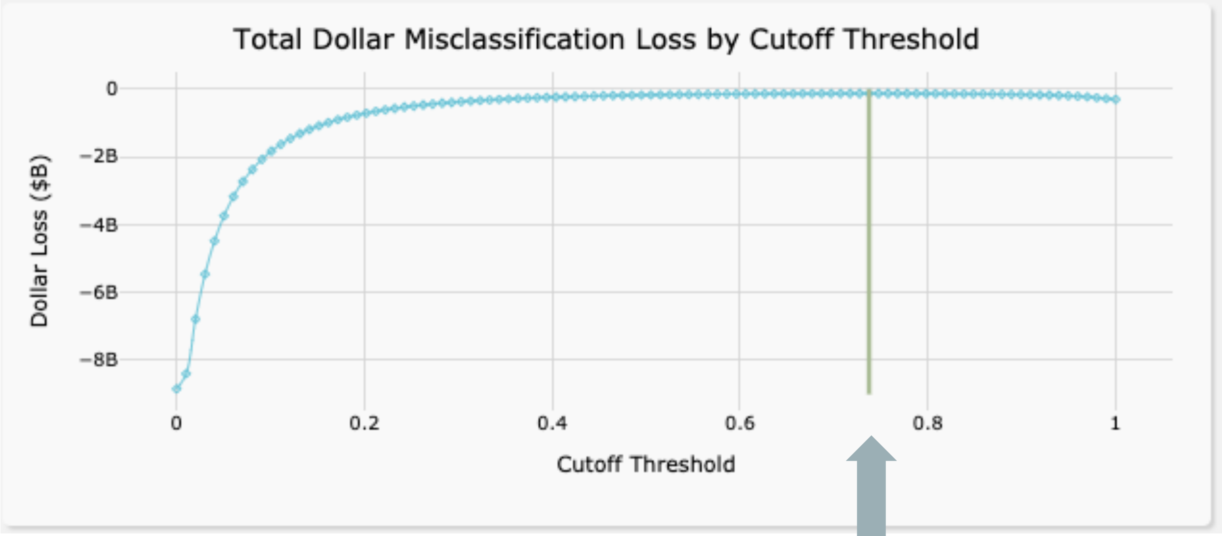
Read more about CatBoost: https://catboost.ai/docs

3D-Modeling CatBoost Prediction Space



Tuning CatBoost Cutoff Threshold

Capturing Dollar Cost of Model Misclassifications



Cutoff value of **0.7374** minimized dollar misclassification loss

Why does high cutoff threshold make sense in this classification setting?

Principal loss of defaulted loans (=0) that were predicted as good (=1) should be expected to outweigh the lost interest income of non-defaulted loans (=1) that were predicted as bad (=0) in the aggregate

Portfolio Optimization Results

Capital Allocation Rules

Simple Rule: For 2018 loan set, leverage CatBoost predictions to allocate capital to loans predicted as ‘good’ and to deny investment for loans predicted as ‘bad’

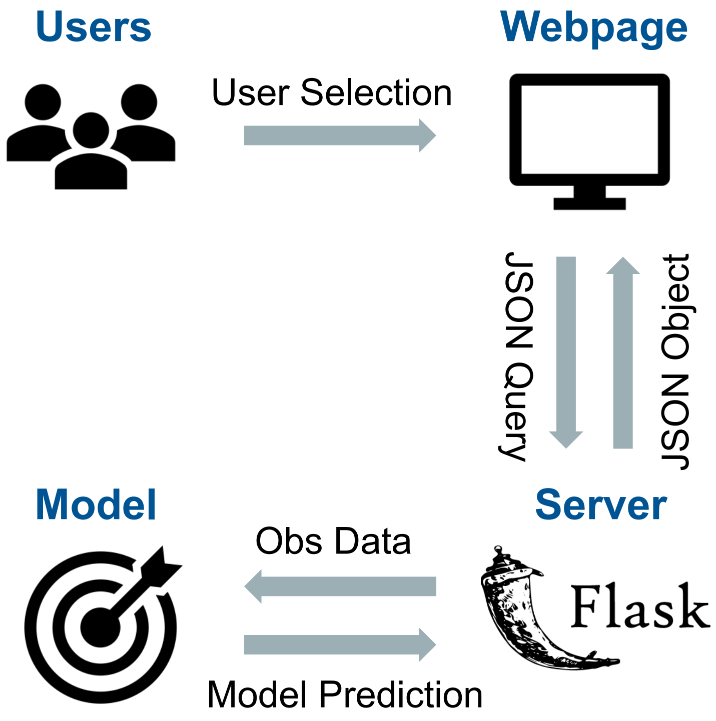
Portfolio Description	36mo loans IRR	Δ vs. Catboost	60mo loans IRR	Δ vs. Catboost
Catboost Portfolio	7.40%	0.00%	10.63%	0.00%
LendingClub Historical IRR[1]	6.30%	-1.10%	8.11%	-2.52%
Baseline Model[2]	5.89%	-1.51%	9.67%	-0.96%

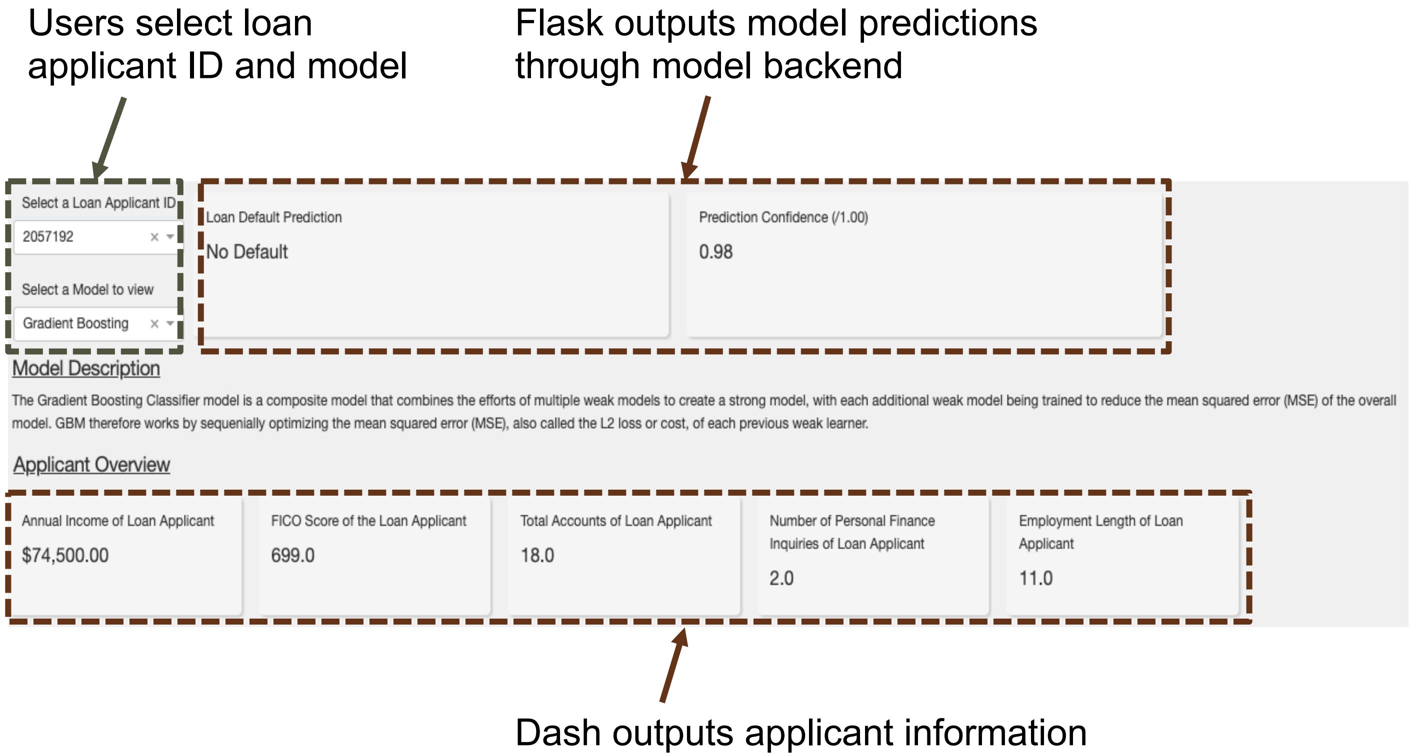
two-sample t-test of our model portfolio against the baseline portfolio further shows these results are **statistically significant to the 1% level**, and that our model produces **1.51%** and **0.96%** of alpha for 36-month and 60-month loans respectively versus the baseline model portfolio

[1] LendingClub IRR inclusive of actual loan recoveries post-default
[2] ‘Baseline’ model predicting all loans as ‘good’ loans indiscriminately

Real-time Machine Learning Predictions using Flask

Goal: Build real-time loan default prediction tools for investors to use





Thanks for tuning in!

You can read more about this project at:

Dash App
http://ec2-3-15-44-208.us-east-2.compute.amazonaws.com/

Blog Post
https://nycdatascience.com/blog/student-works/predicting-loan-defaults-using-machine-learning-classification-models/

Github
https://github.com/philippe-heitzmann/philippe-heitzmann-capstone-app

Contact
If you would have any questions about this project or would like to discuss something else, please feel free to contact me at phil062497@gmail.com

Linkedin
https://www.linkedin.com/in/philippe-heitzmann/

NYC DATA SCIENCE ACADEMY