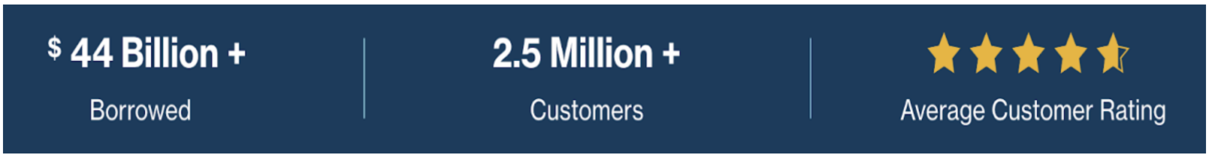
LendingClub

Introduction

Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California. Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee. Lending Club enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request. All loans are unsecured personal loans and can be between $1,000 - $40,000. On the basis of the borrower's credit score, credit history, desired loan amount and the borrower's debt-to-income ratio, Lending Club determines whether the borrower is credit worthy and assigns to its approved loans a credit grade that determines payable interest rate and fees.

As of June 30, 2015, the average Lending Club borrower has a FICO score of 699, 17.7% debt-to-income ratio (excluding mortgage), 16.2 years of credit history, $73,945 of personal income and takes out an average loan of $14,553 that s/he uses for debt consolidation or for paying off credit card debts. The investors had funded $11,217,348,156 in loans, with $1,911,759,192 coming from Q2 2015. The nominal average interest rate is 14.08%, default rate 3.39%, and an average net annualized return (net of defaults and service fees) of 8.93%. The average returns of investment for Lending Club lenders are between 5.47% and 10.22%, with 23 straight quarters of positive returns as of the second quarter of 2013. The statistics on Lending Club's website state that, as of December 31, 2016, 62.3 percent of borrowers report using their loans to refinance other loans or pay credit card debt.



$ 44 Billion +
Borrowed

2.5 Million +
Customers

★★★★⯪
Average Customer Rating

Understanding the Client

Out of the ten fictional characters interested in working with lending club, our client is Taz, the borrower who holds a good credit.

Taz's goal is :

- Get a full funded loan from lending club
- Lowest possible interest rate
- Get a desired loan duration

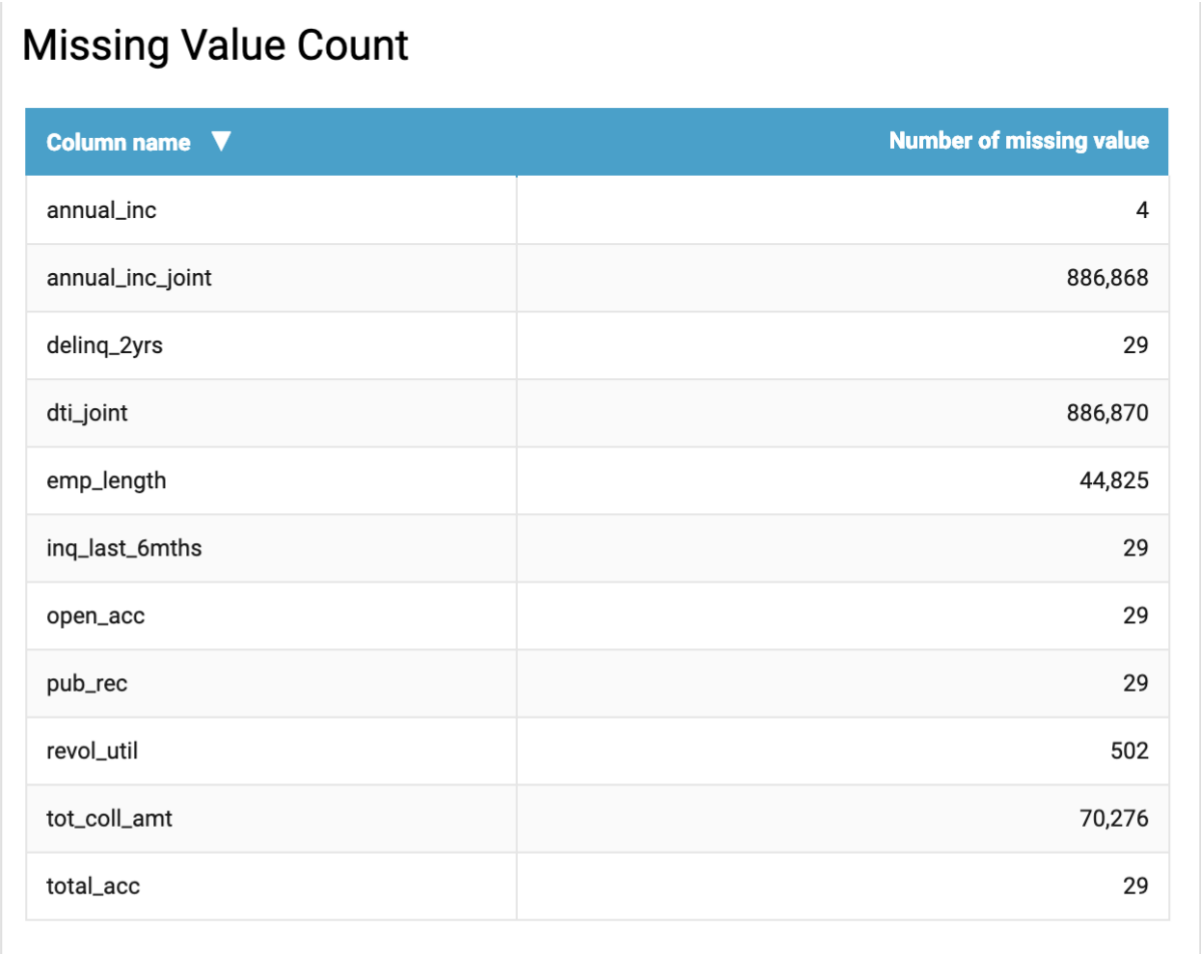
Taz wants us to analyse the lending club dataset to achieve his goals.

Data Quality and Cleaning

For data cleansing, we followed the following steps:

1. After exploring the data, we dropped certain column redundant columns not useful for the analysis.
2. Raw data is inconsistent, so appropriate transformations were applied. Missing values were replaced with either the mean values of the column or 0's, depending upon the column features.
3. The categorical columns were remapped.
4. Data was standardized into numerical format, so that it can be passed to the classifiers in further stages.

Below table shows the count for missing values for different columns.

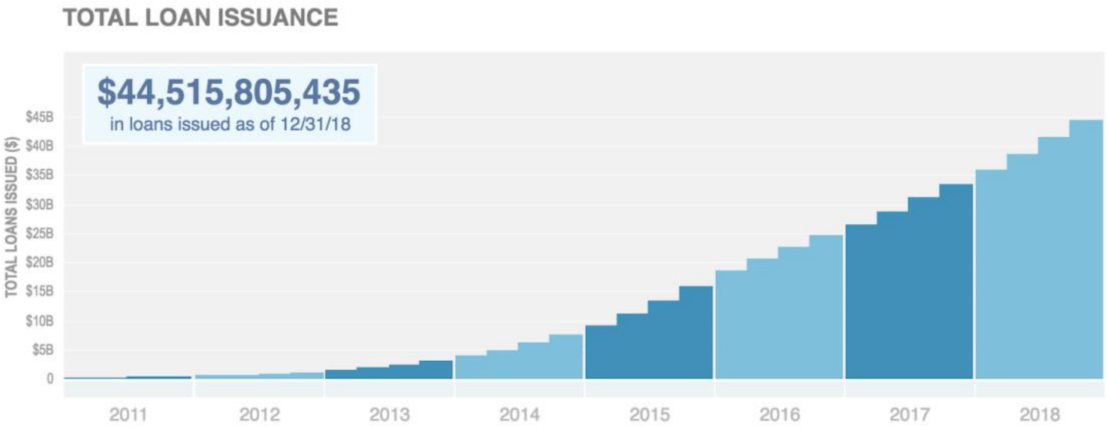


Missing Value Count

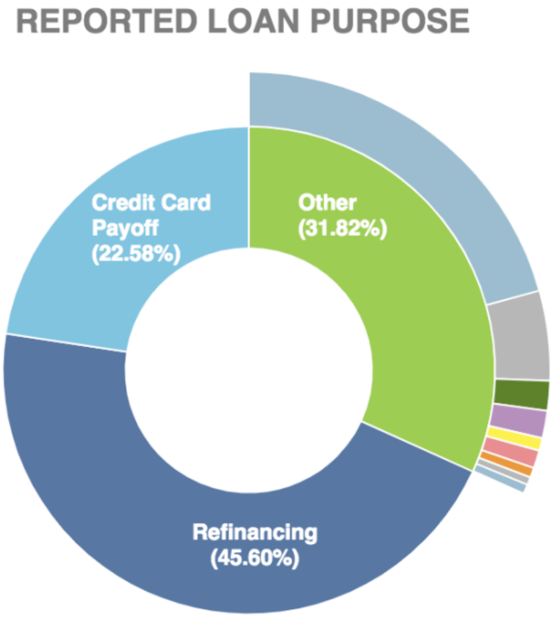
Column name ▼	Number of missing value
annual_inc	4
annual_inc_joint	886,868
delinq_2yrs	29
dti_joint	886,870
emp_length	44,825
inq_last_6mths	29
open_acc	29
pub_rec	29
revol_util	502
tot_coll_amt	70,276
total_acc	29

Lending Club Platform Analysis

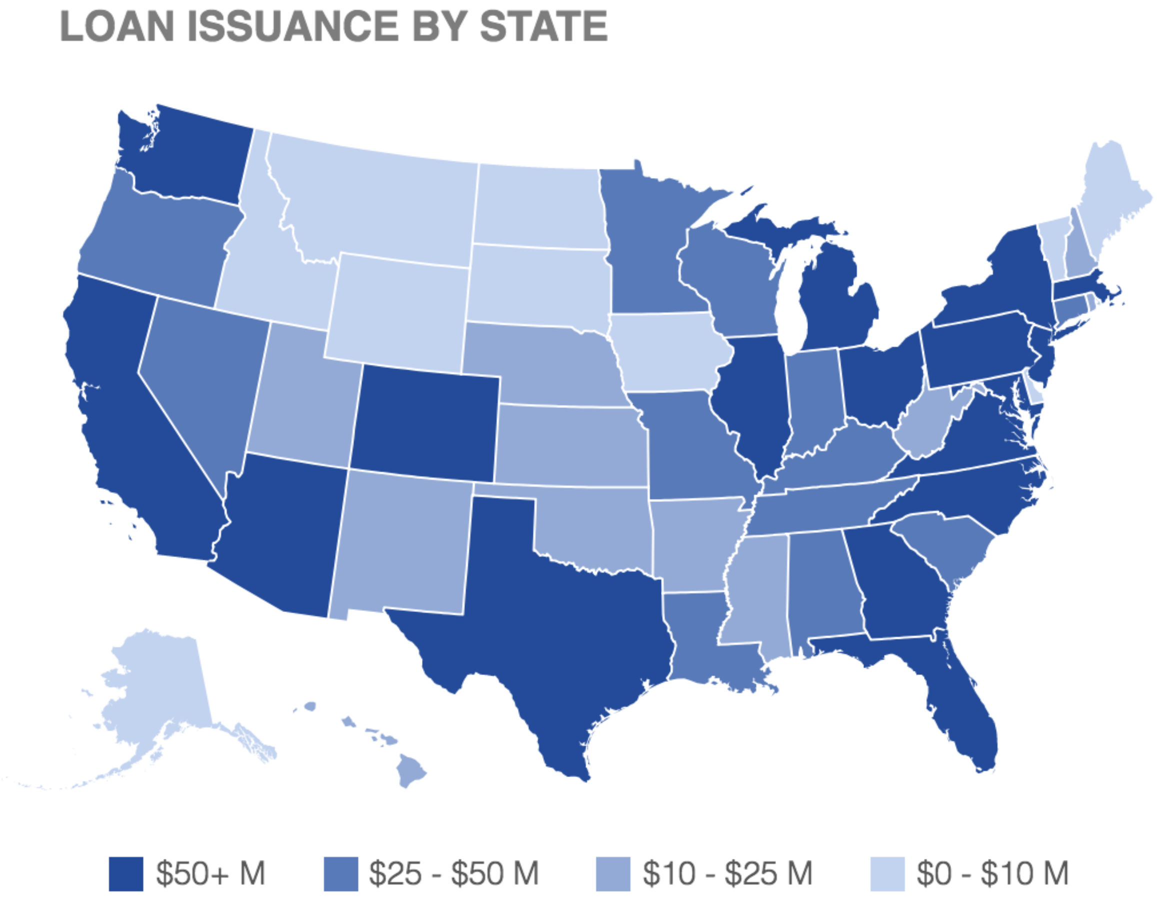
Interest rates are usually set by an intermediary platform on the basis of analyzing the borrower's credit (using features such as FICO score, employment status, annual income, debt-to-income ratio, number of open credit lines).



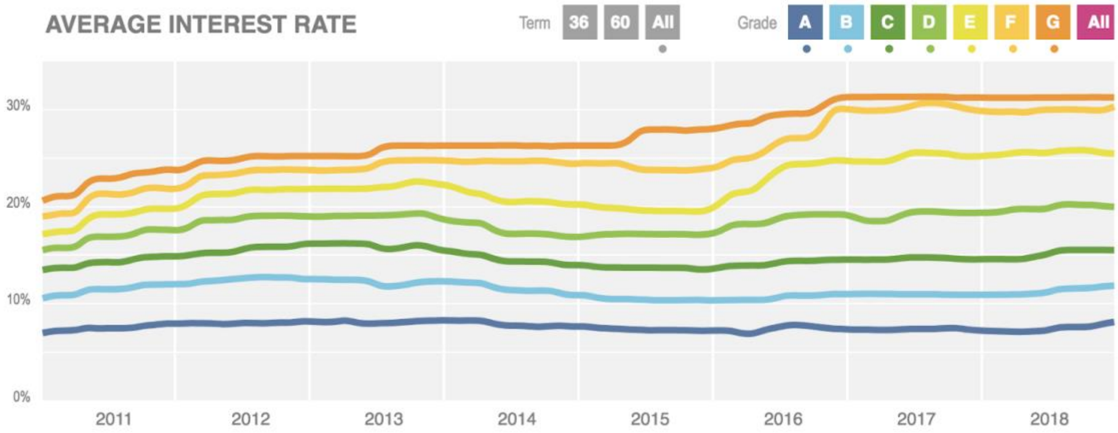
- The number of loans issued by lending club has increased drastically from 2011 to 2018. The number of loans issued increased by 105% from 2015 to 2016, by 46% from 2016 to 2017 and by 35% from 2017 to 2018.



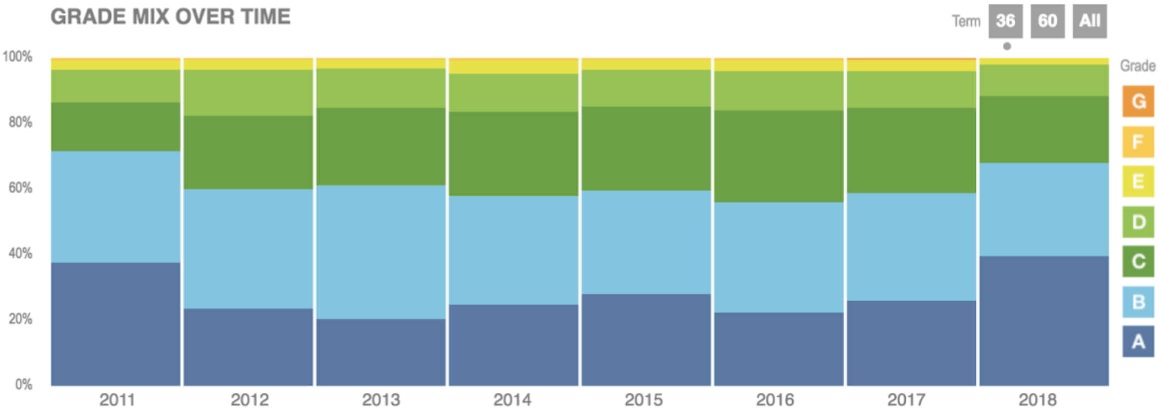
- 68.18% of the loans issued by Lending Club has been used by borrowers to refinance existing loans or to pay off their credit cards bills. 31.82% of the loans issued have been used for other purposes such as vacations, moving and relocation, home buying, medical expenses, business, major purchase, car financing, home improvement etc.



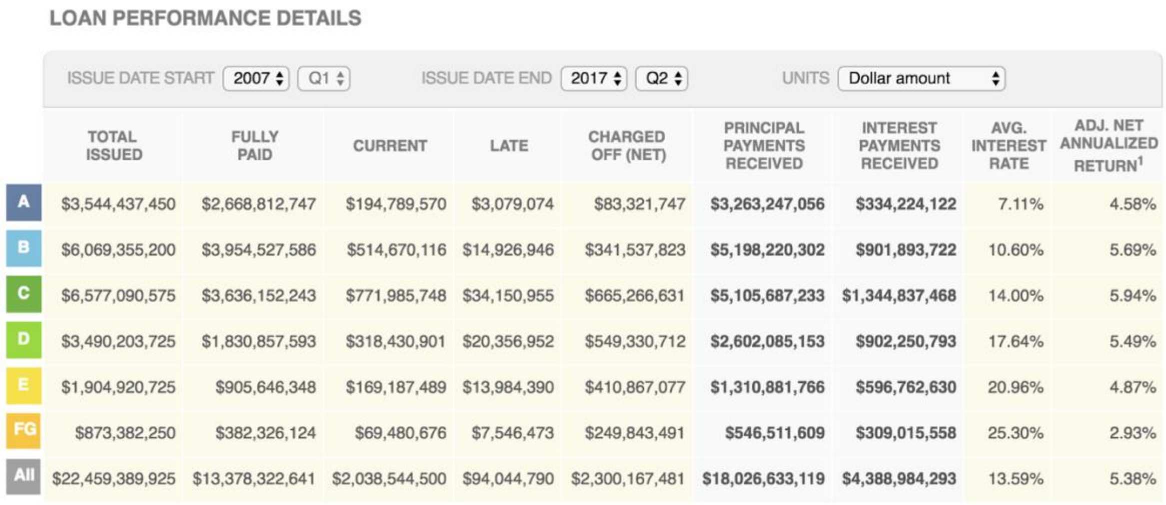
- Lending Club has invested $50+M in the states of Washington, California, Colorado, Arizona, Texas, Florida, Georgia, North Carolina, Virginia, Illinois, Michigan, Ohio, Pennsylvania, Maryland, New York and Massachusetts with the highest being in the state of California with $407,012,861. Lending Club’s investments has been less than $10M in the states of Idaho, Montana, Wyoming, South Dakota, North Dakota, Iowa, Vermont and Maine with the least being in the state of Iowa with $284,258.



- The average rate of interest has been almost constant for grades A, B, C and D. For grades E, F and G the average rate of interest has increased over the years.



- The maximum number of loans were granted to grade B from years 2011 to 2013 and to grade C from 2014 to 2017. But in 2018 maximum number of loans granted changed from grade C to grade B.



LOAN PERFORMANCE DETAILS

ISSUE DATE START 2007 Q1 ISSUE DATE END 2017 Q2 UNITS Dollar amount

	TOTAL ISSUED	FULLY PAID	CURRENT	LATE	CHARGED OFF (NET)	PRINCIPAL PAYMENTS RECEIVED	INTEREST PAYMENTS RECEIVED	AVG. INTEREST RATE	ADJ. NET ANNUALIZED RETURN[1]
A	$3,544,437,450	$2,668,812,747	$194,789,570	$3,079,074	$83,321,747	**$3,263,247,056**	**$334,224,122**	7.11%	4.58%
B	$6,069,355,200	$3,954,527,586	$514,670,116	$14,926,946	$341,537,823	**$5,198,220,302**	**$901,893,722**	10.60%	5.69%
C	$6,577,090,575	$3,636,152,243	$771,985,748	$34,150,955	$665,266,631	**$5,105,687,233**	**$1,344,837,468**	14.00%	5.94%
D	$3,490,203,725	$1,830,857,593	$318,430,901	$20,356,952	$549,330,712	**$2,602,085,153**	**$902,250,793**	17.64%	5.49%
E	$1,904,920,725	$905,646,348	$169,187,489	$13,984,390	$410,867,077	**$1,310,881,766**	**$596,762,630**	20.96%	4.87%
FG	$873,382,250	$382,326,124	$69,480,676	$7,546,473	$249,843,491	**$546,511,609**	**$309,015,558**	25.30%	2.93%
All	$22,459,389,925	$13,378,322,641	$2,038,544,500	$94,044,790	$2,300,167,481	**$18,026,633,119**	**$4,388,984,293**	13.59%	5.38%

- The average interest rate has been highest for grades F and G in lending club. The net annualized return has been highest in the case of grade C followed by grade B and D. The net annualized return has been almost the same for grades B, C and D.

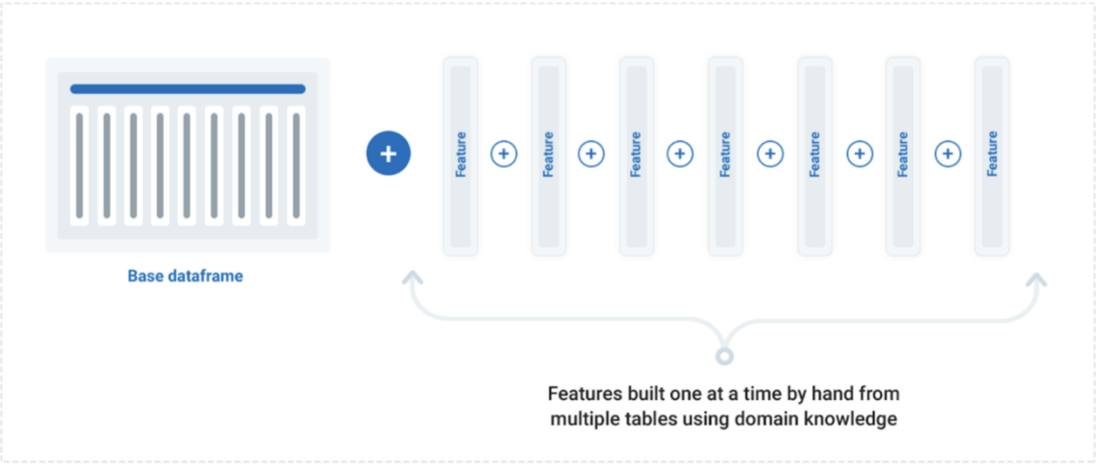
Feature Engineering

Feature engineering is the process of taking a dataset and constructing explanatory variables and features, that can be used to train a machine learning model for a prediction problem. Often, data is spread across multiple tables and must be gathered into a single table with rows containing the observations and features in the columns.

Manual Feature Engineering

The traditional approach to feature engineering is to build features one at a time using domain knowledge, a tedious, time-consuming, and error-prone process known as manual feature engineering. The code for manual feature engineering is problem-dependent.

Manual Feature Engineering



Features built one at a time by hand from multiple tables using domain knowledge

The features derived manually based on the dataset are:

1. borrower_lending_club_satisfaction= funded_amnt/ loan_amnt
2. lending_club_to_investor_satisfaction= funded_amnt_inv/ loan_amnt
3. installment_to_salary_ratio=installment*12/ annual_inc
4. dti (debt to income ratio)
5. payment_received= total_pymnt/ tot_coll_amt
6. credit_utilization=total_debt / all_util

Besides being tedious and time-consuming, manual feature engineering is:

- **Problem-specific:** All of the code I wrote over many hours cannot be applied to any other problem

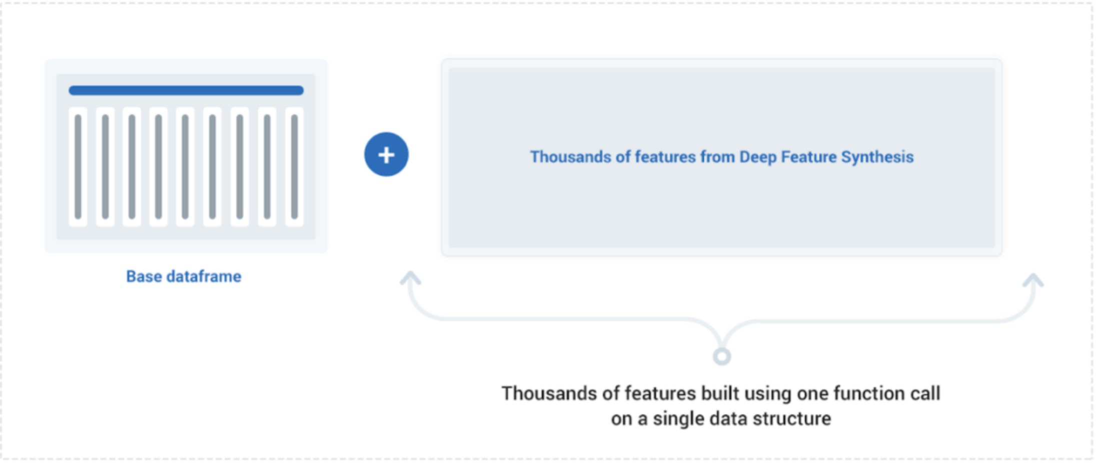
- **Error-prone:** Each line of code is another opportunity to make a mistake

- **Limited Creativity:** The final manual engineered features are limited both by human creativity and patience

Automated Feature Engineering

Automated feature engineering allows even a domain novice such as myself to create thousands of relevant features from a set of related data tables. All we need to know is the basic structure of our tables and the relationships between them which we track in a single data structure called an entity set. Once we have an entity set, using a method called Deep Feature Synthesis (DFS), we’re able to build thousands of features in one function call.

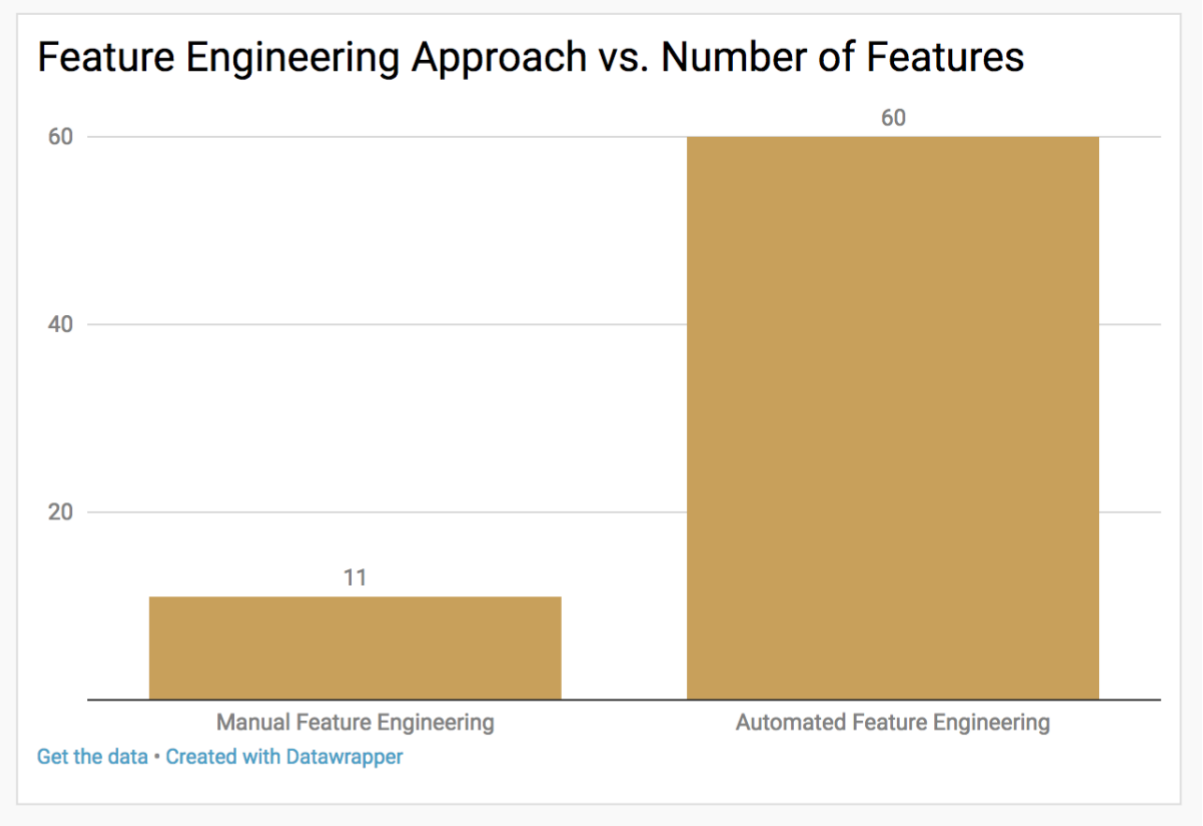
Featuretools Automated Feature Engineering

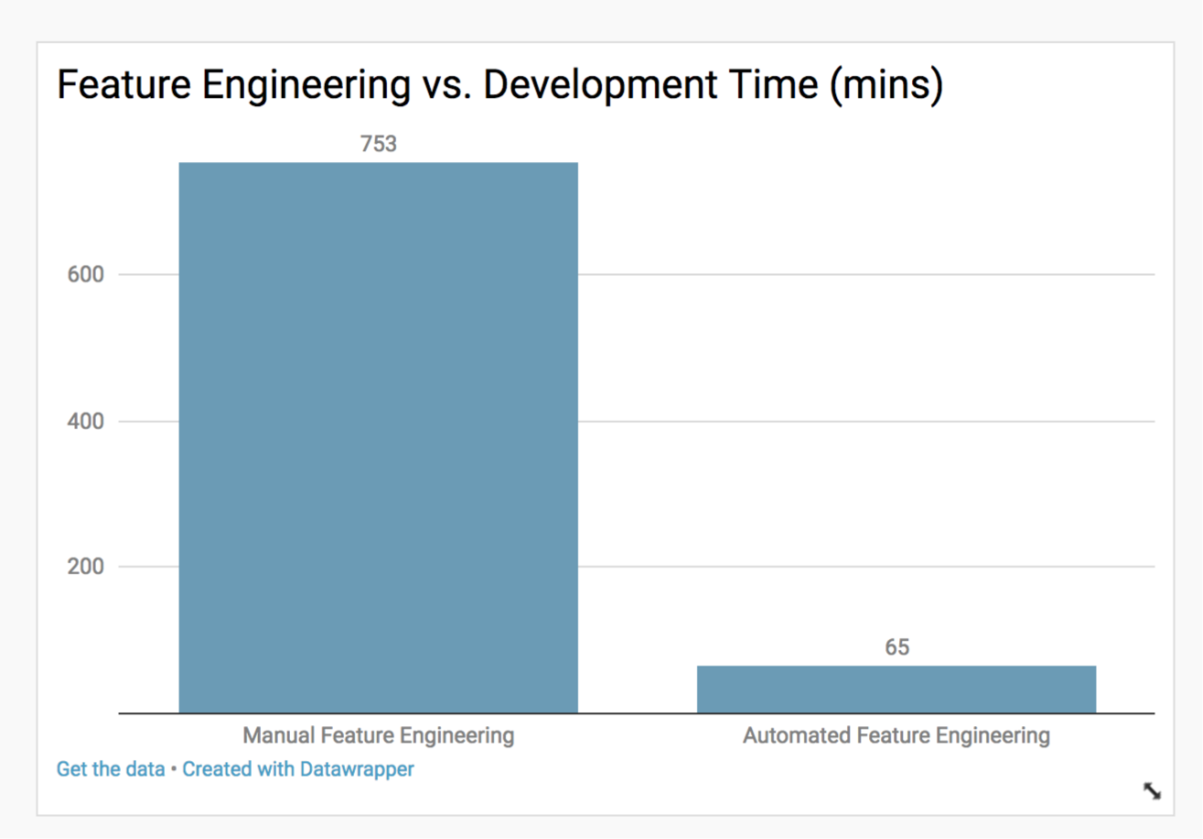

Base dataframe
+
Thousands of features from Deep Feature Synthesis
Thousands of features built using one function call
on a single data structure

Manual vs. Automated Feature Engineering

The main conclusions are that:

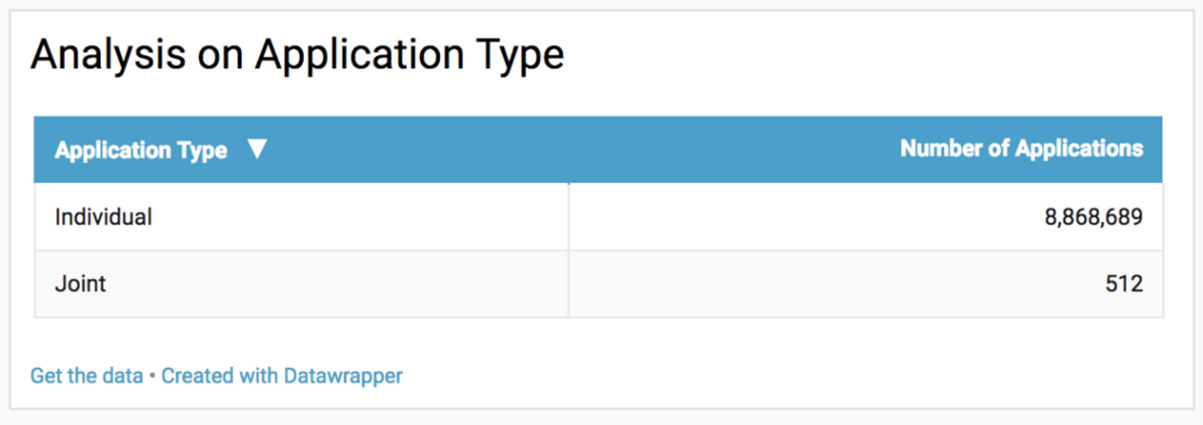
- Automated Feature Engineering reduced implementation time by up to 10x compared to manual Feature Engineering
- Both achieved modeling performance at the same level or better level
- Automated Feature Engineering delivered more number of interpretable features with real-world significance compared to manual one.


Feature Engineering Approach vs. Number of Features
60
40
20
11
60
Manual Feature Engineering
Automated Feature Engineering
Get the data • Created with Datawrapper


Feature Engineering vs. Development Time (mins)
753
600
400
200
65
Manual Feature Engineering
Automated Feature Engineering
Get the data • Created with Datawrapper

Analysis based on various parameters

1. The total number of individual applications are 1700 times more than the joint ones.

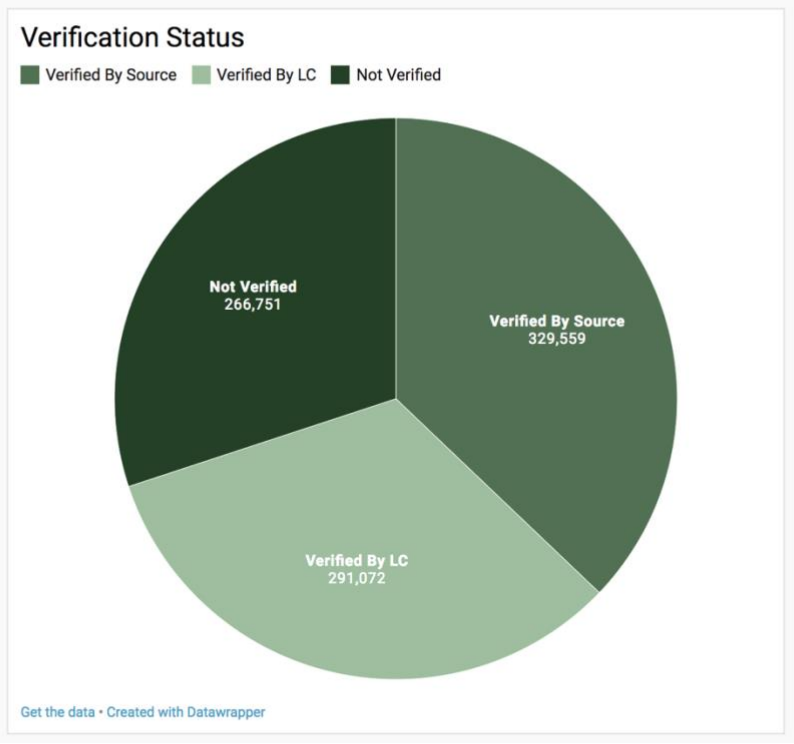


Analysis on Application Type

Application Type ▼	Number of Applications
Individual	8,868,689
Joint	512

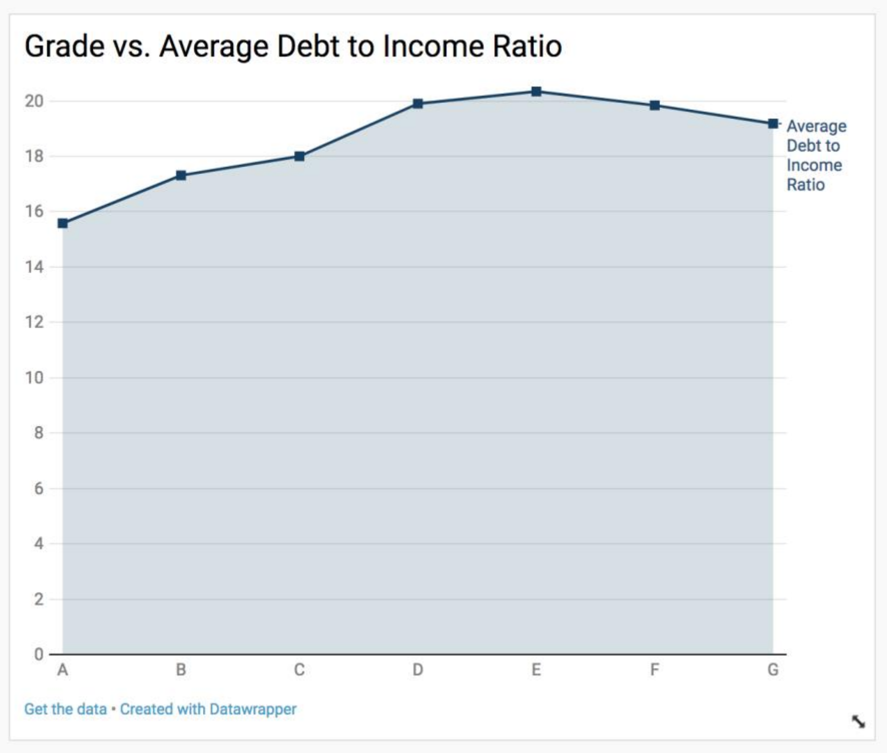
Get the data • Created with Datawrapper

2. For approximately 28% percent of total applicants, income is not verified



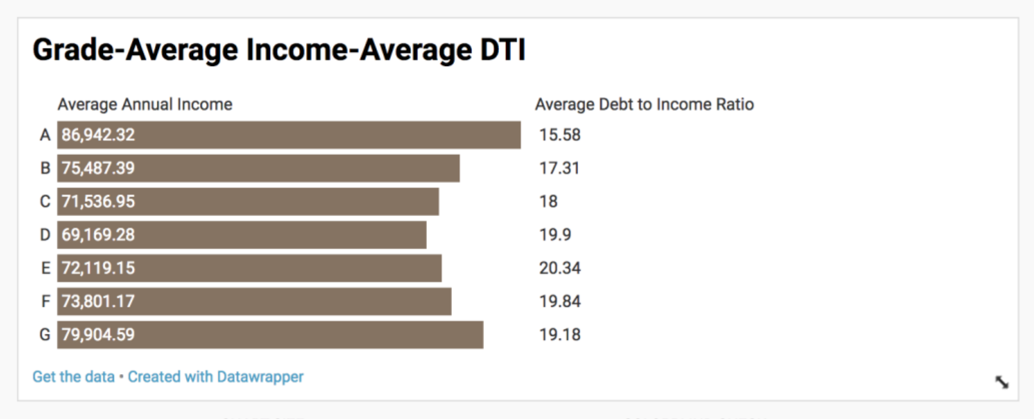
Get the data • Created with Datawrapper

3. The average annual income of grade G applicants is higher that of grade C but the debt to income ratio is higher in case of grade G. Average rate of interest of grade G applicants is 25.63 which highest and implies that high debt to income ratio can lead to higher rate of interests.



Grade vs. Average Debt to Income Ratio

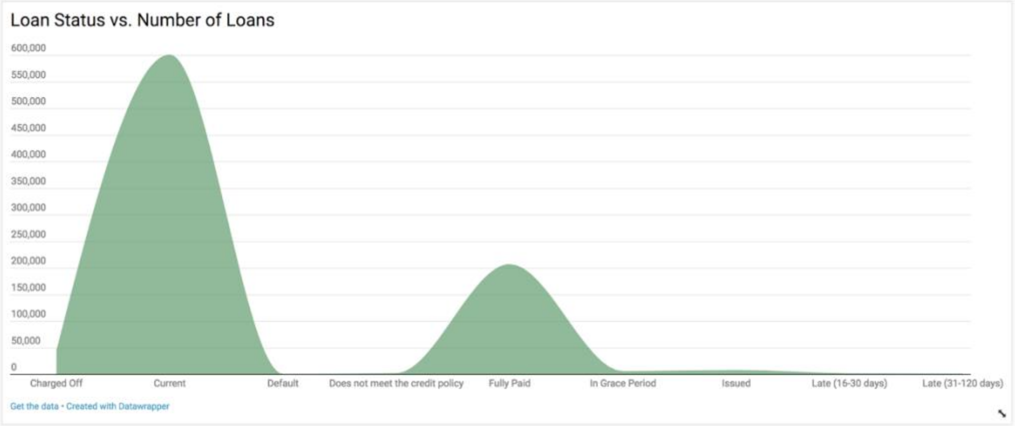
Get the data • Created with Datawrapper



Grade-Average Income-Average DTI

	Average Annual Income	Average Debt to Income Ratio
A	86,942.32	15.58
B	75,487.39	17.31
C	71,536.95	18
D	69,169.28	19.9
E	72,119.15	20.34
F	73,801.17	19.84
G	79,904.59	19.18

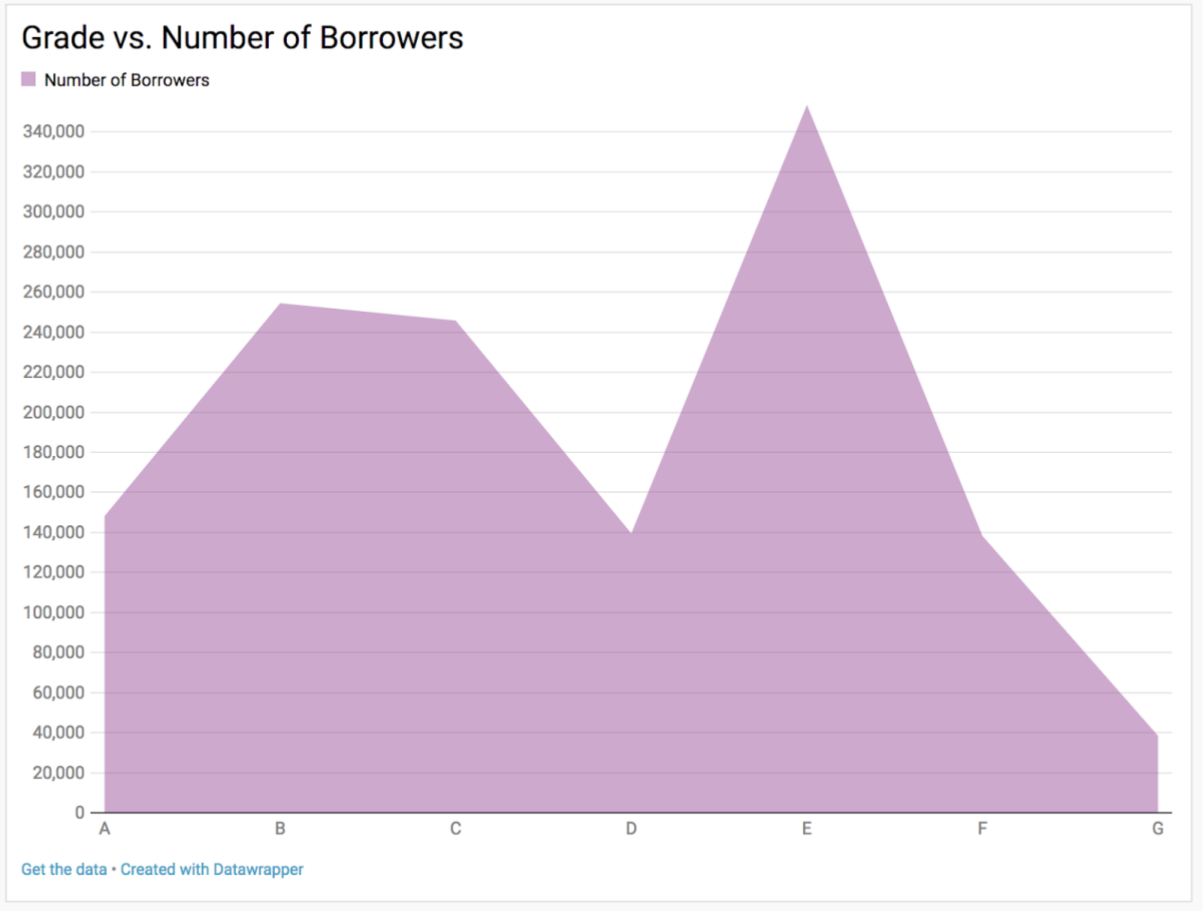
Get the data • Created with Datawrapper

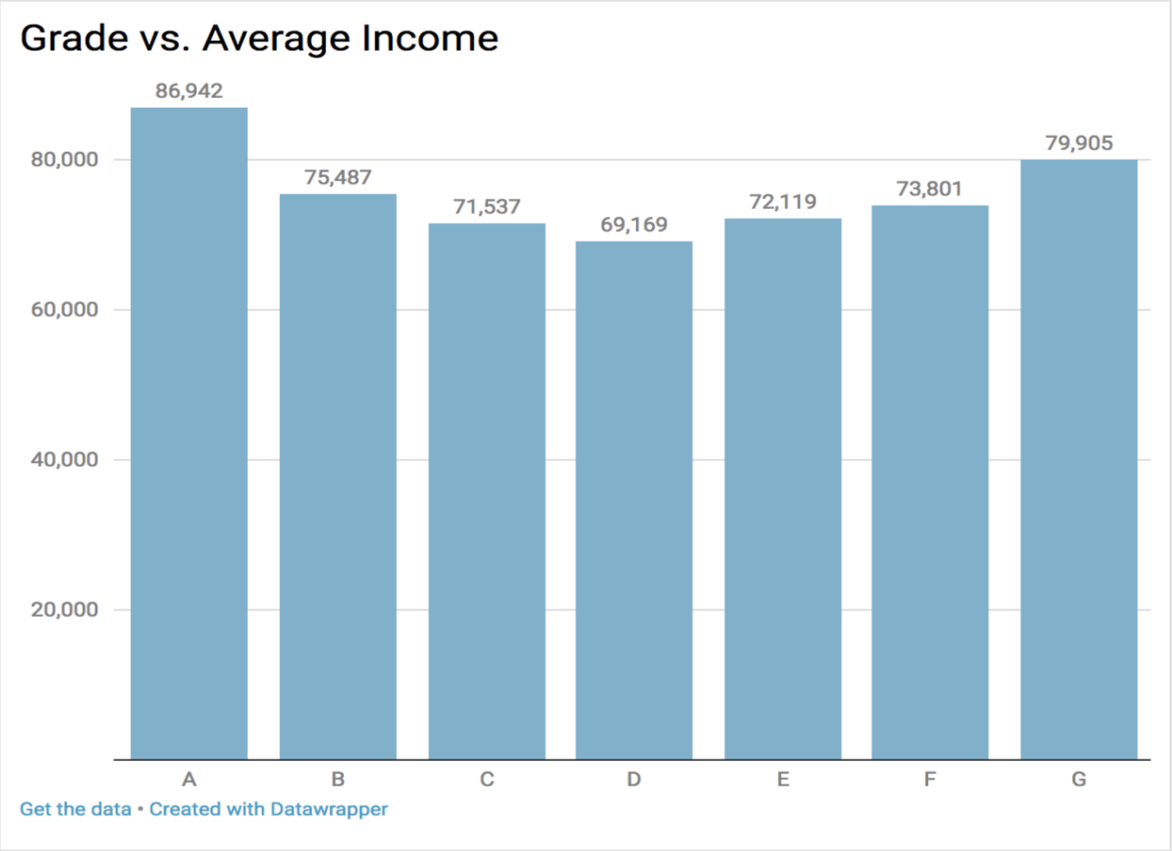


Loan Status vs. Number of Loans

Get the data • Created with Datawrapper

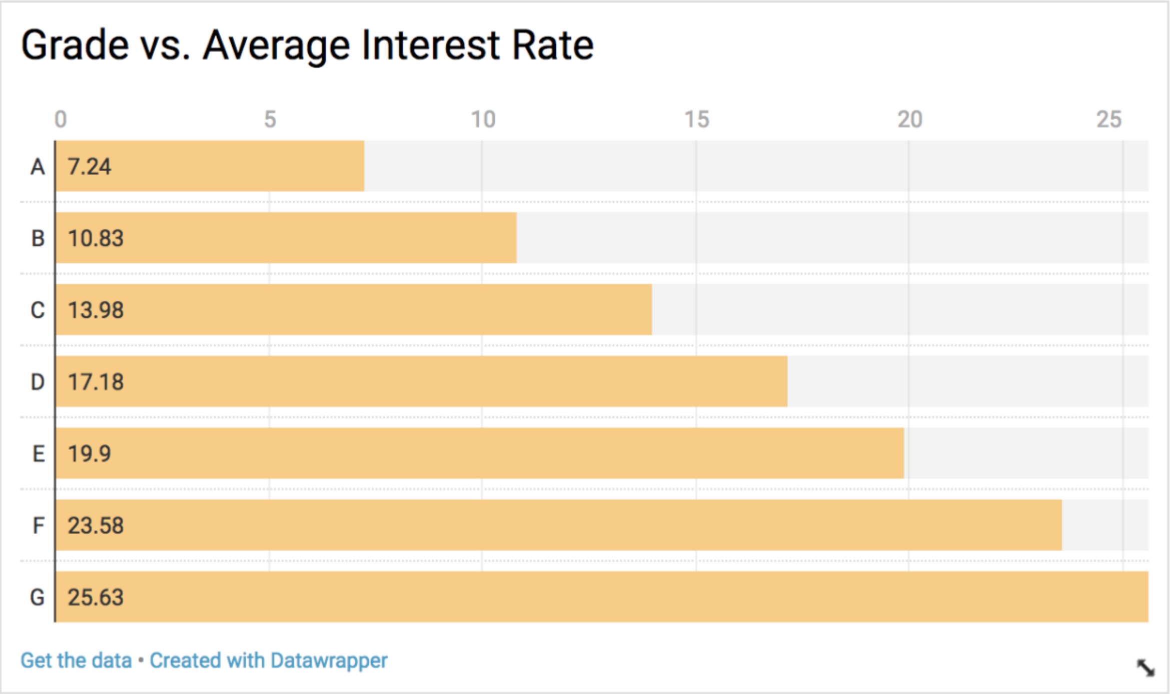
4. Highest number of borrowers belong to Grade E with average dti of 20.34 and an average income of 72,119.15 USD





Grade vs. Average Income

Get the data • Created with Datawrapper



Grade vs. Average Interest Rate

Get the data • Created with Datawrapper

Model Performance (Regression, Random forest, Neural Networks)

Design Specifications

The dataset was divided into training(60%), validation(20%) and testing datasets(20%).

Models used:
1. Linear Regression
2. Random Forest Regressor
3. Neural Network

The training and validation sets were then passed to each of the classifier.

To train your model : clf.fit(x,y)

This is a regression problem. Hence, we cannot use classification accuracy metrics.

There are different ways to calculate the error here :

1. Mean_square_error
2. The 'score' function (which calculates the R^2) provided by each SKLearn Model.

MSE should be as close as possible to 0.0 and 'score' result is like the accuracy, so, it should be close or equal to 1.
Usually 'score' function is not used. There are various different metric functions that can be tried.

Based on the scores obtained, Linear Regression is the best fit model for Lending Club analysis.

K-fold cross validation.

The inbuilt cross validation function does not support different metric functions hence, we need to write a custom function.
The dataset for this function is train and test only.

Hyperparameter Tuning

Hyperparameters contain the data that govern the training process itself.

Effect of Hyperparameter Tuning

Every model is simply a math equation with various coefficients that can be tuned to fit a specific dataset. clf = LinearRegression() means the equation has been initialised with default values for those coefficients. We have to tune these coefficient values and find the ones that are a right fit for us. Finding the right values for a specific dataset is still a topic that is in research. It's just like finding the right weights for your neural network. Let's say 'a' is a coefficient in the equation for LinearRegression(), now usually you can set the value between a specific range (provided in documentation), when you fit your dataset in the model, it will not fit well with the default value for 'a'. So what you do is try all the possible values for 'a' and compare the result. There will be some values for which the data might overfit or underfit. But, there will be a value that will be the best fit. But, models have several parameters that can be tuned. GridSearch provides is a script to generate all the combinations of these coefficient values.

Say you have,

```
    a = range[1, 5]
    b = [True, False]
    a=1,b=True
    a=2,b=True
    a = 1, b=False
    a = 2, b = False
    and so on.
```

Grid search will help generate all the possible combinations

Next we initialize the model and plug in these parameter values, train and test our model and log the accuracy. Once done with all the combinations, we can compare which ones were best. This helps us find the best parameters for our model and even help improve the accuracy.

Hyper Parameter tuning is really important for getting the best fit for our dataset. It is a tool which helps us select the best features in the dataset.

AutoML runs a series of different models and picks the best one out for you.

AutoML is more human interpretable and reproducible.