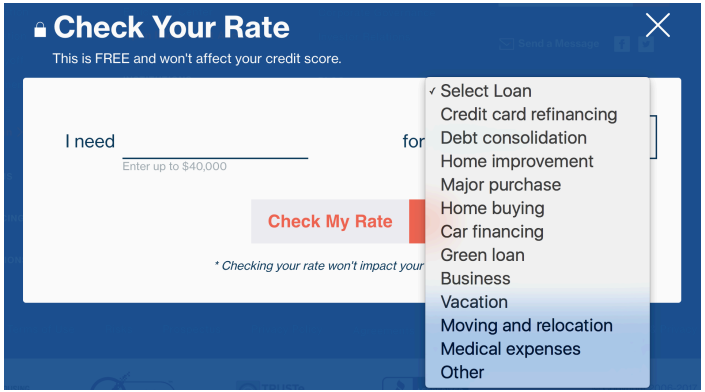
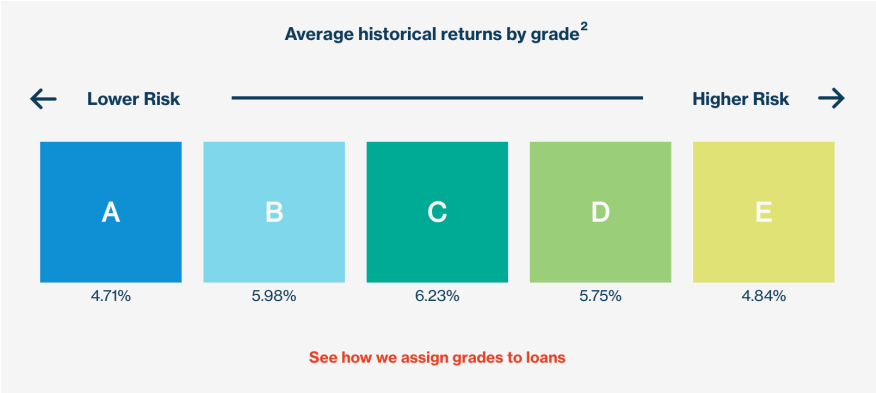
Predicting Good/Bad P2P Loans

Supervised Learning I Binary Classification

Darren K. Lee I GA Winter 2017

P2P Lending Landscape





Ask an interesting Question

“Can we tell if a borrower will be likely to default or not before they are approved for their loan?”

Goals

Minimize investment risk, and reject bad loans with higher probability to default or delay in payment.

Accept better candidates onto the loan platform

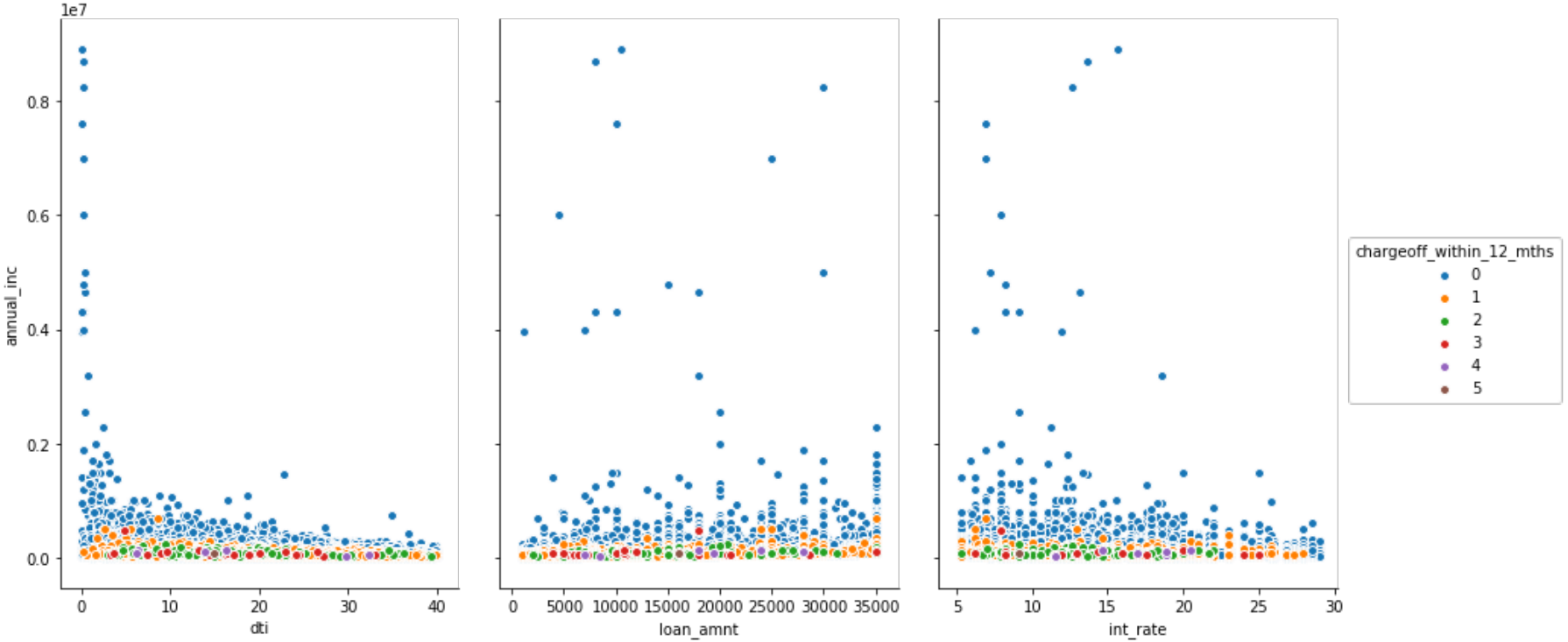
Make the model interpretable, and less of a Black Box model.

Get the Data

CSV → 400,000 Rows → 145 Features

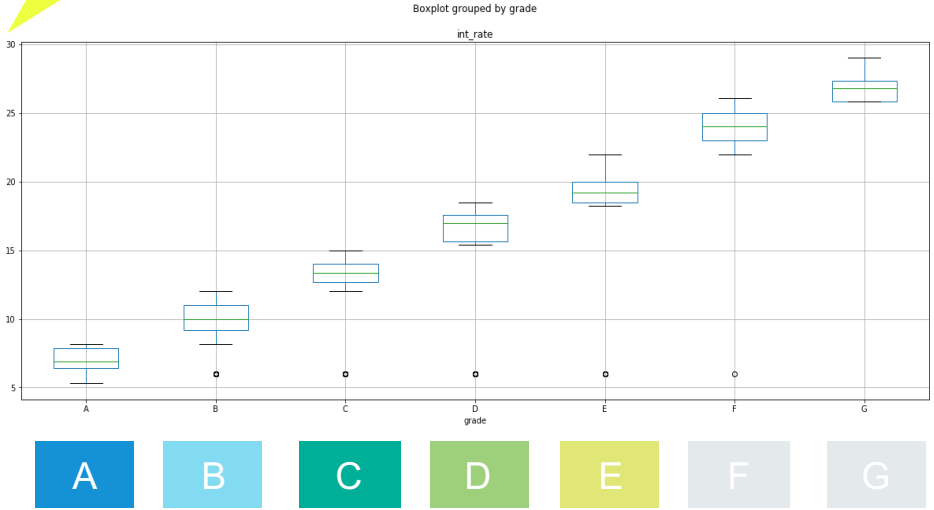
- **2015** dataset chosen because the idea is to gather more completed loans that are either already paid in full or already charged off by now
- **183,660** rows in final dataset
- **85 Features** in the machine learning dataset.

Explore the Data

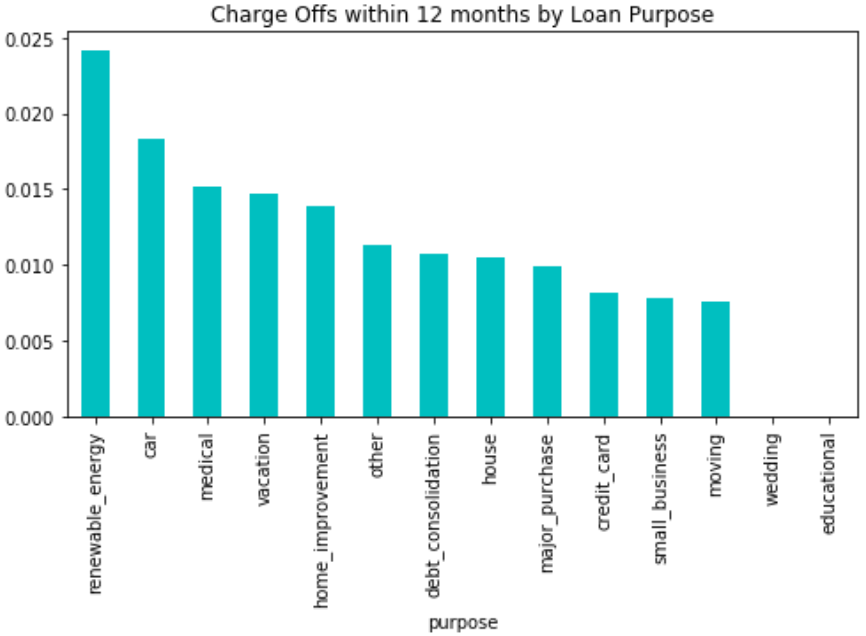


Explore the Data (cont.)

30%?!

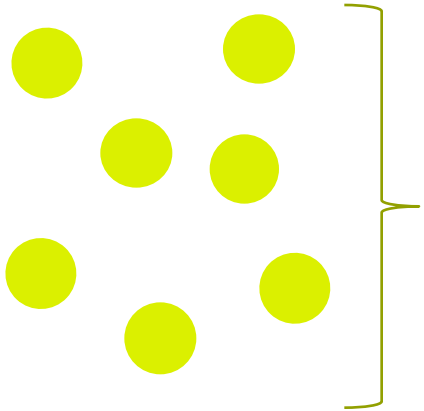


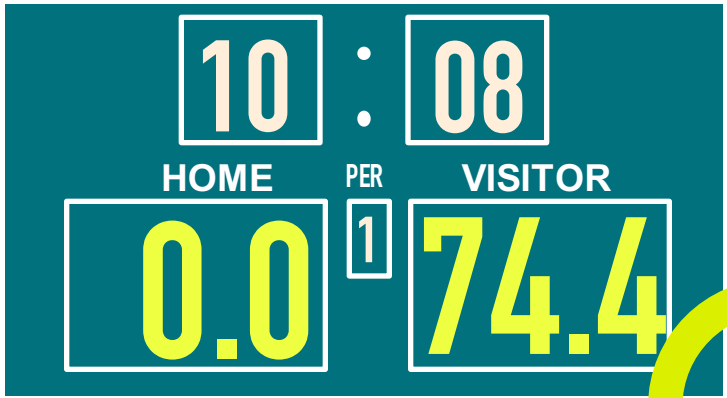
A B C D E F G



Model the Data | Targets

The goal is to transform the **loan_status** column into a *binary feature*. There are 7 types of loan statuses, and the target varables kept in the model will be:





Good Loan = “Fully Paid”

Bad Loan = “Charged Off”

What do we mean by the wrong prediction?

Amortization

$$A = P\frac{r(1+r)^n}{(1+r)^n - 1}$$

	Predicted Good Loan (0)	Predicted Bad Loan (1)
Actual Good Loan (0)	TN	FP
Actual Bad Loan (1)	FN	TP

- **FP** = We rejected an application that would've been a Good Loan
- **FN** = We accepted an application, and it turned out to be a Bad Loan

Lifetime Value (LTV) for Bad Loan Group

Amortization Schedule Calculator | $ 16000 | Term 4 | % 15 | ZIP 96771 | Jan | 2018 | Calculate

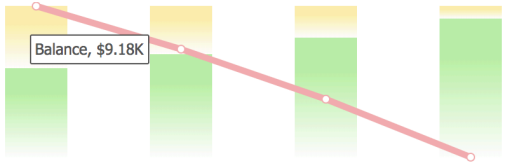
Loan Summary

$445 Monthly Payment

$21,374 Total of 48 Payments

$5,374 Total Interest Paid

Dec, 2021 Pay-off Date



Mortgage Rates | Amortization Schedule

Date	Interest	Principal	Balance
Jan, 2018	$200	$245	$15,755
Feb, 2018	$197	$248	$15,506
Mar, 2018	$194	$251	$15,255
Apr, 2018	$191	$255	$15,000
May, 2018	$188	$258	$14,742
Jun, 2018	$184	$261	$14,481
Jul, 2018	$181	$264	$14,217
Aug, 2018	$178	$268	$13,950

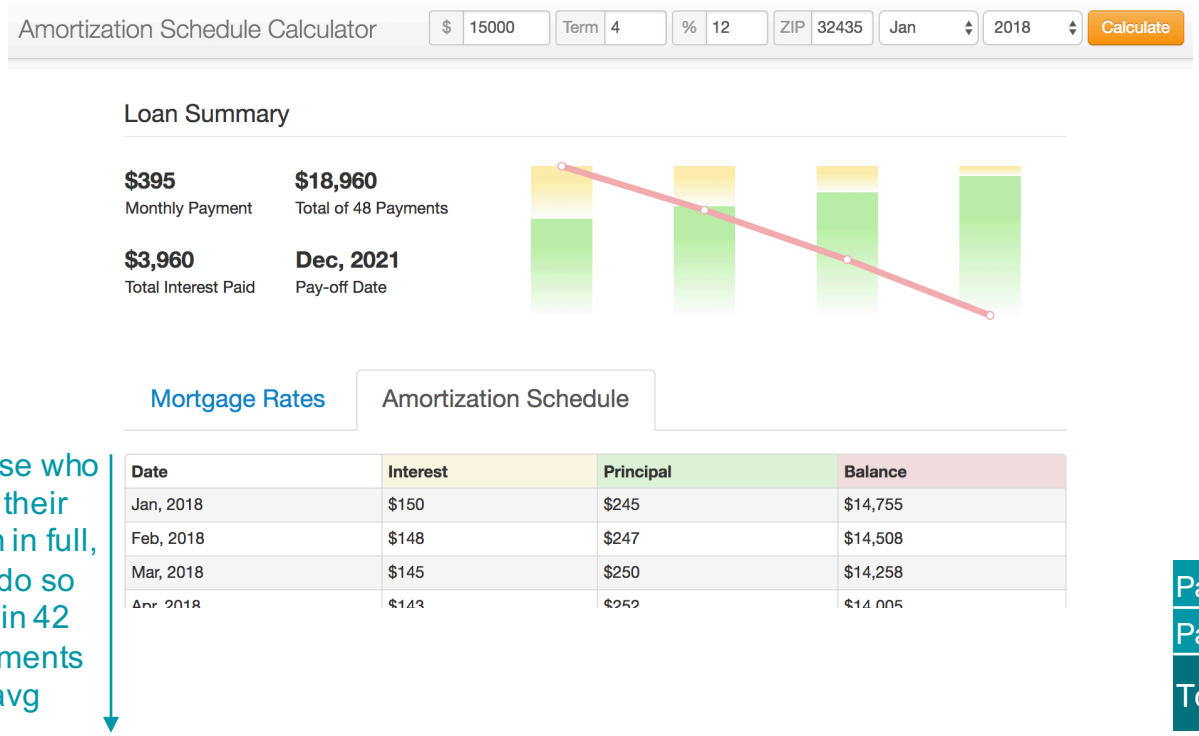
Those who default, will do so within 16 payments into the loan

Averages:

- 48 months term
- $16,000 Loan Amount
- 15% Interest Rate
- ~$7,000 Total payback

Remaining Outstanding Interest	($2,565)
Remaining Outstanding Principal	($11,683)
Total Avg Loan Loss	($14,248)

Lifetime Value (LTV) for Good Loan Group



Those who pay their loan in full, will do so within 42 payments on avg

Averages:

- 42 months term
- $15,000 Loan Amount
- 12% Interest Rate
- ~$17,000 Total payback

Paid Principal	$13,000
Paid Interest	$3,900
Total Avg Profit Loss	$3,900

Feature Engineering | Feature Selection

Decision Tree as our first Off-The-Shelf model to get a list of Feature_Importances_ (with 43 hand picked subset of features to begin with). Also optimized with randomized search.

Used the dataset of 80+ features for Random Forest Classification (also used randomized search)

Next: Transform data Scale/PCA + AdaBoostClassification().

Next: Feature Engineering emp_title using Word2Vec via h2o.ai

(*) Removed from the dataset because they are post application metrics

DT Feature Importance	RF Feature Importance
int_rate	total_pymnt*
term	loan_amnt
avg_cur_bal	funded_amnt
dti	int_rate
revol_util	installment
acc_open_past_24mths	total_rec_int*
mort_acc	term
loan_amnt	avg_cur_bal
annual_inc	dti

Decision Tree & Random Forest Classifier models

Cross Validated ROC_AUC: **70.1%**

Cross Validated Accuracy Score: **75.0%**

RandomizedSearchCV HyperParameters	Decision Tree Classifier	Random Forest Classifier
criterion	entropy	gini
max_depth	8	6
max_features	20	10
min_samples_split	2	10
n_estimators	-	41
Precision Score	52.85%	38.70%
Recall Score	18.32%	67.67%

Cross Validated ROC_AUC: **70.5%**

Cross Validated Accuracy Score: **74.4%**

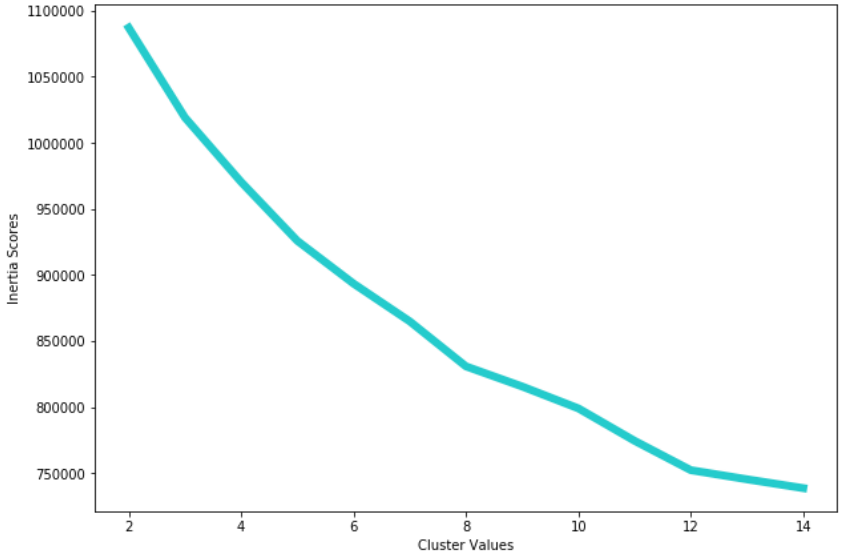
Decision Tree	Predicted Good Loan	Predicted Bad Loan
Actual Good Loan	42,460	2,550
Actual Bad Loan	12,740	2,858

Random Forest	Predicted Good Loan	Predicted Bad Loan
Actual Good Loan	28,297	16,713
Actual Bad Loan	5,043	10,555

emp_title | Word2Vec & K-Means Clustering

```
import h2o
h2o.init()
from h2o.estimators.word2vec import H2OWord2vecEstimator
from h2o.estimators.gbm import H2OGradientBoostingEstimator
```

- **Tokenize** words
- **Word2Vec** will find synonyms then create word vectors for each employment title

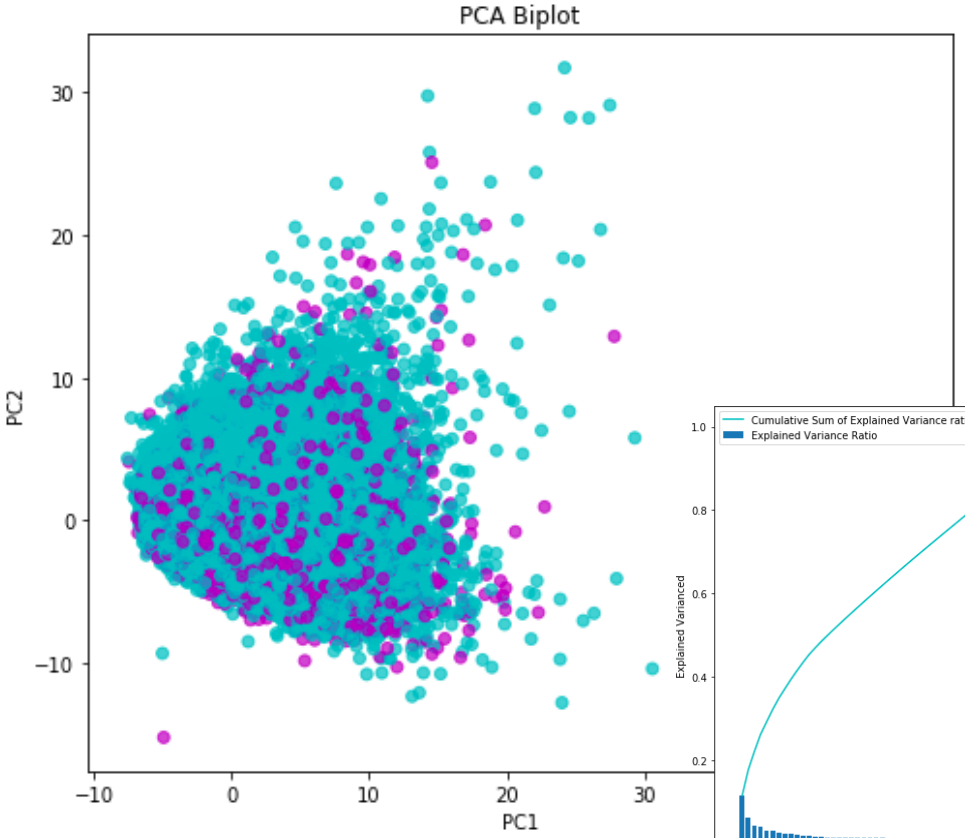


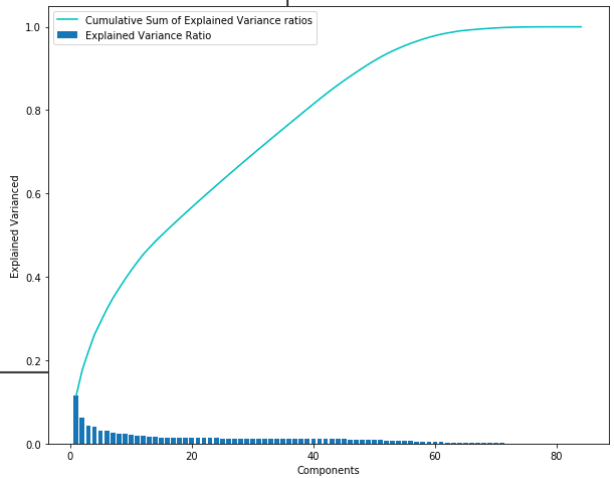
K-Means to cluster vector output

- **Selected 6 Clusters:**
  0. Miscellaneous jobs
  1. Engineer or Technical jobs
  2. Managers & CEO's
  3. Teachers
  4. Hospital jobs
  5. Drivers / operators

- Assigned **new labels**
- Created new pd.DataFrame with new Cluster labels

PCA | Dimensionality Reduction

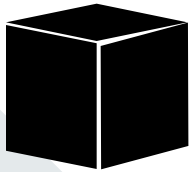




PC #1 Top Weights	PC #2 Top Weights
open_acc	acc_open_past_24mths
chargeoff_within_12_mths	tot_coll_amt
total_rev_hi_lim	tot_cur_bal
total_il_high_credit_limit	pub_rec_bankruptcies
num_rev_accts	funded_amnt

- Dimensionality Reduction from 80+ columns down to 2
- The Biplot with just 2 Principal Components accounts for ~20% explained variance.
- As you can see, it will be difficult to separate classes with this type of overlap

AdaBoostClassifier model



Cross Validated Accuracy Score:
74.4%

AdaBoost CLF on PCA

# of Principal Components	% Cumulative Explained Variance	Train/Test Accuracy Score:
2	~17.5%	~74%
44	~85%	~74%
85	100%	~74%

n_estimators	100
Precision Score	38.7%
Recall Score	64.1%

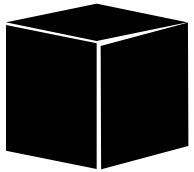
Cross Validated Accuracy Score:
75.7%

CV ROC_AUC Score:
72.95%

AdaBoost CLF	Predicted Good Loan	Predicted Bad Loan
Actual Good Loan	42,556	2,454
Actual Bad Loan	12,480	3,118

n_estimators	260
Precision Score	~56%
Recall Score	~20%

Final Model Evaluation - GBM



GBM Parameters	Values
Max Features	50
Max Depth	6
n_estimators (default)	100

Precision Score	**57%**
Recall Score	21%

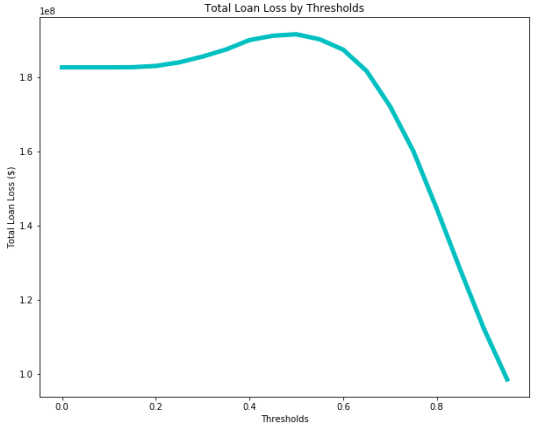
CV Accuracy Score:
75.9%

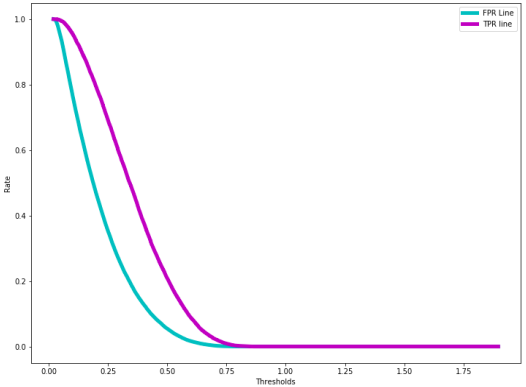
CV ROC_AUC Score:
73.4%

- Performed RandomizedSearchCV with 9 iterations, over 45 different models. This took a very long time!
- The model is *slightly* better than the previous AdaBoost Classifier model. Please note that I am very happy to be beating Lending Club's model as is!
- The loans that do not make this funnel, will have to go through an additional under writing process with Under Writing agents asking the borrower to submit additional information.

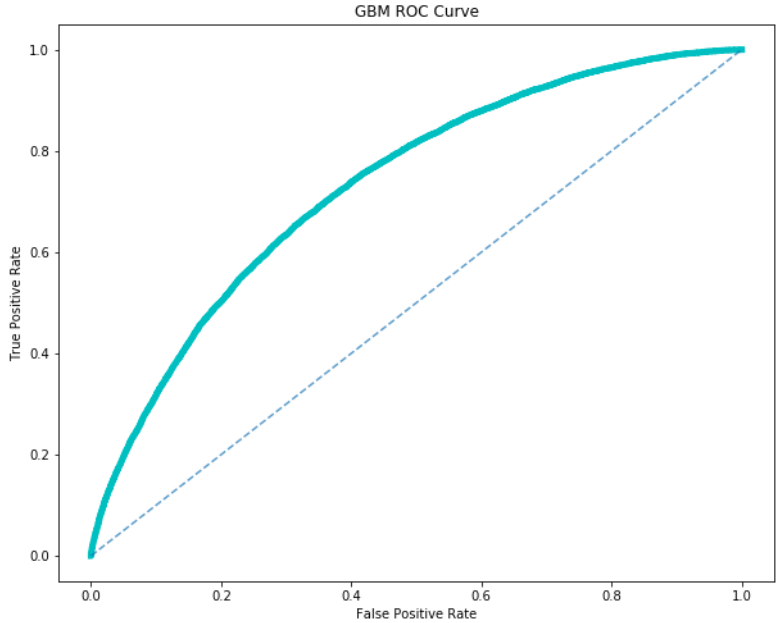
Gradient Boosting Classifier	**Predicted Good Loan**	**Predicted Bad Loan**
Actual Good Loan	**42,552**	**2,458**
Actual Bad Loan	**12,338**	**3,260**

Model Evaluation - GBM





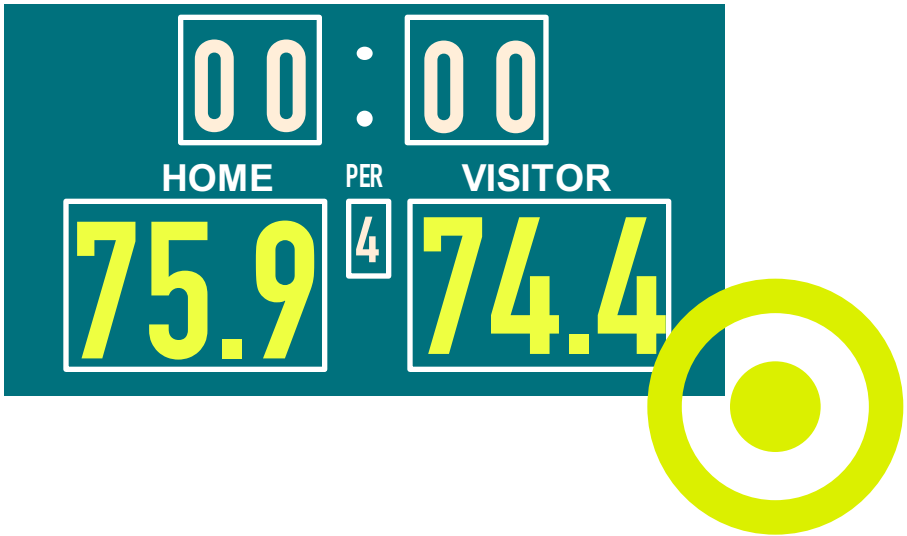
- Threshold = 0.95
- Total Loan Loss = **$95,329,642**
- Avg Total Loan Loss = **$1,573** / Borrower



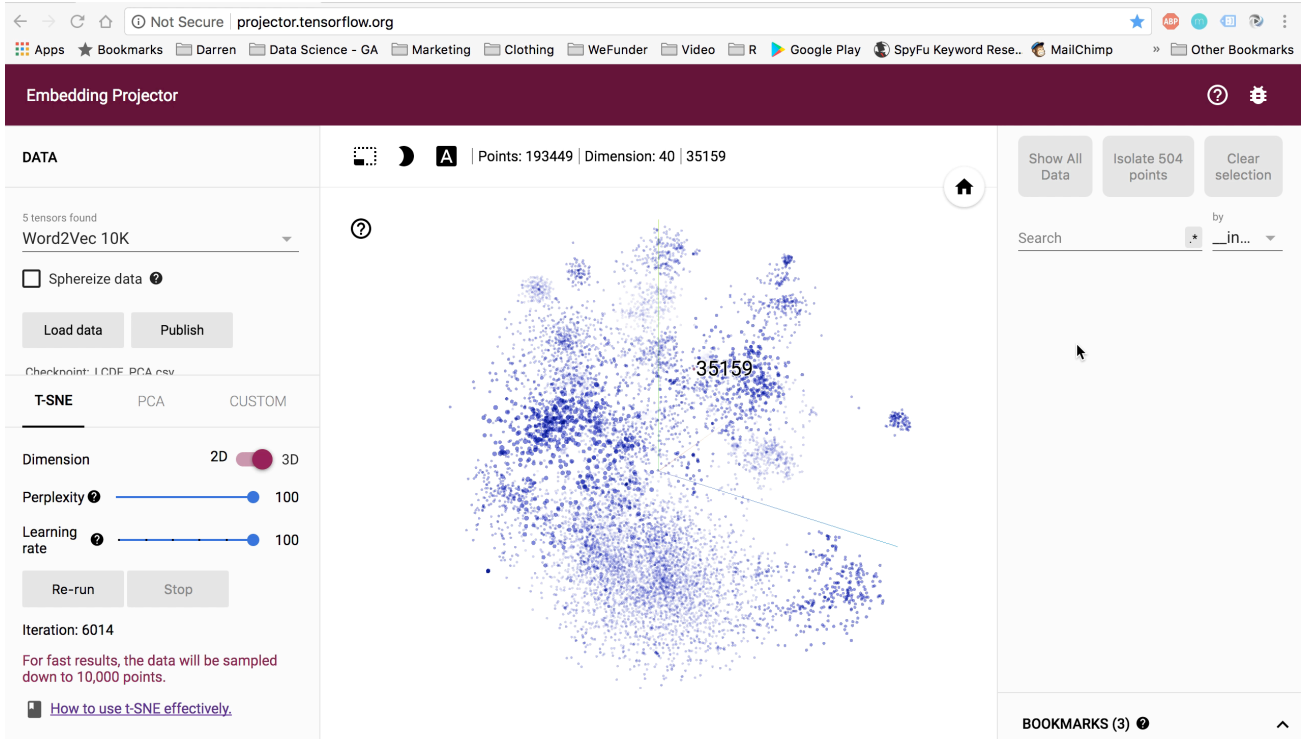
Summary

If I had more time…

- I would work on a multi-class prediction model with the targets being the **grade** (A – E) that Lending Club uses.
- More hyperparameter tuning with Keras Deep Learning & grid search.
- Feature Engineering & sentiment analysis with “desc” field
- Drop insignificant features



EXTRA: TensorFlow Projector



- PCA file with 40 of 80 dimensions plotted using t-SNE
- Uploaded .tsv file
- It's free! Great visualization tool for clusters and high dimensions
- You can cluster responses (using the Index in this example)

Questions?