Analysis of LendingClub Data

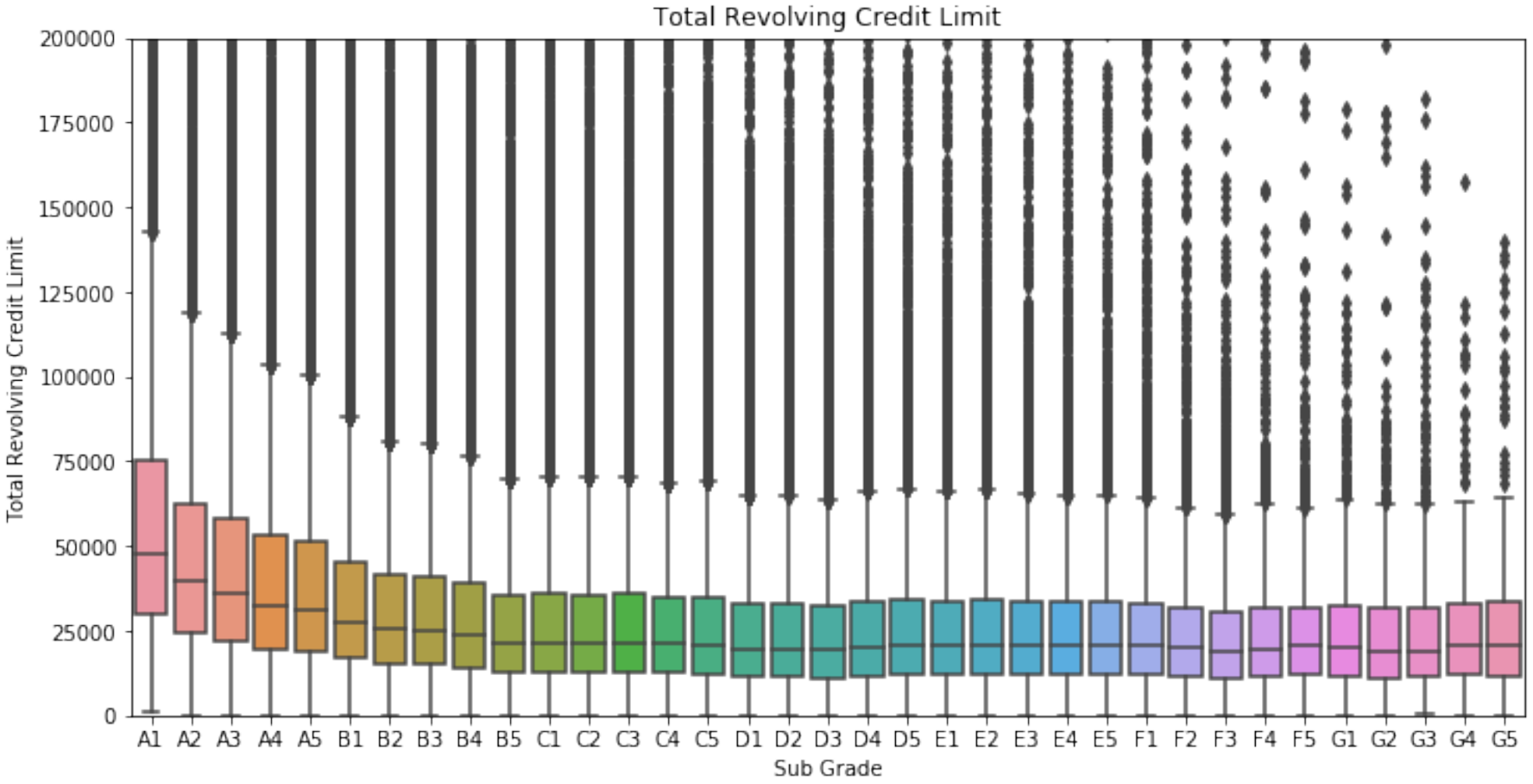
David Ermer

- https://www.kaggle.com/wendykan/lending-club-loan-data/downloads/lending-club-loan-data.zip
- Data on 887379 loans
- FICO scores removed
- LendingClub assigns 35 sub-grades

Statistical Analysis

- Data Columns show significant statistical difference across sub-grades
- However there is a lot of overlap in the distributions
- This will effect predictive modeling

Total Revolving Credit Limit



Predictive Modeling

- Classification using
  - SVM
    - linear and rbf kernels
  - KNeighborsClassifier
- Accuracy ~0.80
  - Confusion matrices
    - Results not adequate to apply to selection process

Confusion Matrix

Testing	Predicted: Good	Predicted: Bad
Actual: Good	47821	2693
Actual: Bad	10278	856

Can Prediction Be Improved?

- Unbalanced classes?
- Data not adequately differentiated?

Unbalanced Classes

- Split data 25%/75% into testing and training sets
- From training set generate 10 sets that include equal numbers of ‘good’ and ‘bad’ loan classes
- Train 10 Classifiers and let them ‘vote’

Results

- KNeighborsClassifier
  - n_neighbors = 12
  - Accuracy 0.7652
- Threshold = 10 (all vote ‘bad’)
- Didn’t solve problem

	Predicted: Good	Predicted: Bad
Actual: Good	44971	5541
Actual: Bad	9945	1191

Add Uncertain Classification

- Two voting thresholds
  - $T_1 = 8$, $T_2 = 10$
- KNeighborsClassifier from before
- Accuracy = 0.7663
  - ‘uncertain’ not counted as a prediction

	Predicted: Good	Predicted: Bad	Uncertain
Actual: Good	39833	5174	5505
Actual: Bad	7536	1838	1762

Conclusion

- Data as analyzed has two much overlap between the classes to generate a useful predictive model