How to prevent another financial crisis on Wall Street:

Predicting the riskiness of real estate loans

Katelyn Gao[1], Susanne Halstead[2]

1 Introduction

Starting in Fall 2008, there was a meltdown on Wall Street that contributed to what some say was the worst economic recession since the Great Depression. One of the major culprits was Commercial Mortgage Backed Securities (CMBS). CMBS are securities that are backed by loans on commercial properties, such as hotels and office buildings.

CMBS are traded in markets, so we may use the interest rate as a measure of the riskiness of the asset. However, since the interest rate is also determined by other factors such as loan maturity, we instead use loan spread as a proxy for riskiness.

The goal of our project was to predict the average spread of a loan given its other features and analyze which features have greater effects on its riskiness. In addition, given aggregate data on the number of defaults by property type, we examined whether those property types with higher default rates also had higher average spreads. The hope is that we will be able to predict the riskiness and default likelihood of a loan using only its features, and gain insight into business practices and market conditions.

The paper is organized as follows. Section 2 gives some background on CMBS and explains relevant terminology. Section 3 describes our data set and how we pre-processed it. Section 4 explains the data mining methods we used, and Section 5 gives our results. Section 6 concludes.

2 Background

CMBS are usually grouped by property type, and the five main categories are Office, Multifamily, Retail, Lodging, and Mixed Use. In practice, the property type plays a large role in risk assessment; Lodging is considered the riskiest and Multifamily the safest.

Other factors that affect risk evaluation in practice are Debt service coverage ratio and Loan-to-Value, possibly nonlinearly and dependent on the property type. Next, we define some loan features used in this paper.

- Originator: the financial entity that processes the loan, usually the lender or broker
- Spread: the interest rate, adjusted for market rates
- Loan-to-Value (LTV): ratio of loan amount to property value
- Debt service coverage ratio (DSCR): ratio of cash available for debt servicing to the value of payments
- Net operating income (NOI): annual revenues minus expenses for the property
- Amortization type: method by which the loan principal decreases over time
- Debt Yield: ratio of NOI to loan amount
- Cap rate: ratio of NOI to property value

3 Data

3.1 Description

Our main data set contained information on real estate loans originated in New York and California in 2012-13 from Trepp [1]. Each observation/loan included the features listed in Section 2, as well as other features of the loan and underlying property, such as maturity date, benchmark, balance, occupancy, and appraised value.

A secondary dataset lists aggregate information on delinquency and defaults by property type.

[1] kxgao@stanford.edu
[2] susanne2@stanford.edu

3.2 Pre-processing

The loans in our primary data set are benchmarked by several different baseline interest rates: 10-year treasury bills, 7-year treasury bills, 5-year treasury bills, and LIBOR. To make them comparable, we removed the loans benchmarked by LIBOR, added 40 basis points to the spread of loans benchmarked by 5-year treasury bills, and added 20 basis points to the spread of loans benchmarked by 7-year treasury bills.

In addition to the features listed in Section 2, we also included the following features:

- County (where the property is located)
- Property type
- Loan purpose: Refinance, Acquisition, or Recapitalization
- Occupancy

There was a good deal of missing data. If an observation was missing Loan purpose or Originator, we deleted it. If it was missing Debt yield, DSCR, LTV, or Cap rate, we imputed the missing value with the mean over all observations. If it didn't have occupancy, we imputed the missing value with the mean occupancy for all loans with the same property type.

Instead of NOI, we used NOI per square feet or unit. Each observation either included information on the number of units in the property or the square footage. If an observation was missing NOI per square feet or unit and had square footage, we imputed the missing value with the mean NOI per square feet or unit over all observations with square footage present. Likewise, if an observation was missing NOI per square feet or unit and had the number of units, we imputed the missing value with the mean NOI per square feet or unit over all observations with the number of units present. Finally, we normalized NOI per square feet or unit by whether the observation included information on square footage or number of units.

4 Methodology

We carried out separate analyses for New York and California, and then compared the results.

Several of the features in our data set may be correlated. For example, DSCR and Debt Yield are essentially measuring the same thing; DSCR was used traditionally and Debt Yield is a more modern metric. In addition, as stated in Section 2, the effect of certain features on spread may be nonlinear. They may also depend on the property type. Therefore, our primary analysis took into account correlated features and the possibility of interaction effects.

4.1 LASSO and Forward Stepwise

Lasso and stepwise regression are methods that build multiple linear regression models on the data while also performing feature selection [2]. Forward stepwise regression is a greedy algorithm that starts with the intercept and adds features that create a best fit model at each step. Backwards regression starts with a model on the full feature set and then removes features with the least influence on the fit. Hybrid models consider either forward or backwards steps in each iteration. We used an implementation of the stepwise algorithm that considers the AIC measure to evaluate the various options.

LASSO is a method that considers all features simultaneously [2], [4]. It builds a model that minimizes the residual sum of squares while also constraining the absolute value of the sum of the coefficients; thus forcing coefficients of low-influence features to 0.

These methods will help us gauge which features are most influential on average spread and will build models that consider these important features.

4.2 SVMs and Trees

Next, we explored the possibility of interactions in the effects of certain features on spread. To do so, we first did Support Vector Machine regressions with the ε-insensitive error measure[3] using a variety of kernels, including radial and

[3] As explained in the Elements of Statistical Learning Section 12.3.6, we may extend the SVM analysis to regression by solving a L^2-penalized regression problem.

polynomial [2]. Our motivation was that, in addition to building a predictive model, by finding out which kernel led to the lowest error, we'd be able to discover what types of interactions are present in the data.

After discovering that there were indeed interaction effects in the model, we built a predictive model using regression trees. Regression trees fit piecewise constant models on the feature space [2], and thus take into account possible interactions. We pruned the trees to minimize the cross-validated mean squared error (MSE).

For each tree, we also computed the relative importance of variables [3]. Proposed by Breiman et al. (1984), the importance of a variable is equal to the total decrease in MSE over nodes when that variable is used as the splitting variable while growing the tree.

To improve upon regression trees, we also built random forests on the data. Random forests, by bootstrapping the data, growing small trees to each bootstrap sample, and averaging over all samples, are able to reduce the variance of our predictions while keeping the bias small [2]. As with regression trees, we computed the relative importance of variables using the same metric.

4.3 Aggregate Defaults

The dataset on delinquency and defaults gives an overview of actual delinquencies and defaults by property type, thus providing an ex-post assessment of how risky properties are. We compare the ex-post ranking of riskiness with the mean spread in each category.

5 Results

5.1 LASSO and Forward Stepwise

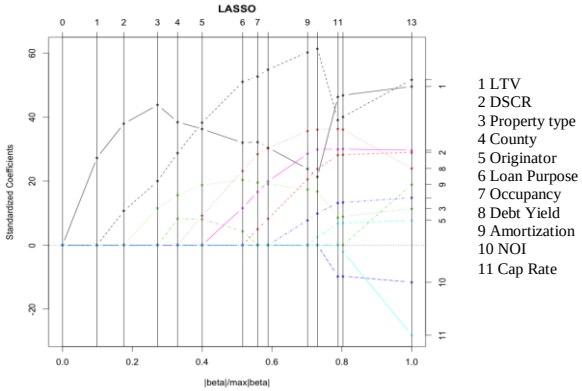
Lasso and stepwise produced results in similar ranges of MSE of each other:

MSE	New York	California
Stepwise	3663.793	1708.547
Lasso	3929.154	1756.799

We proceed to finding the important features selected by these models.

5.1.1 New York

The following figure is the LASSO path.



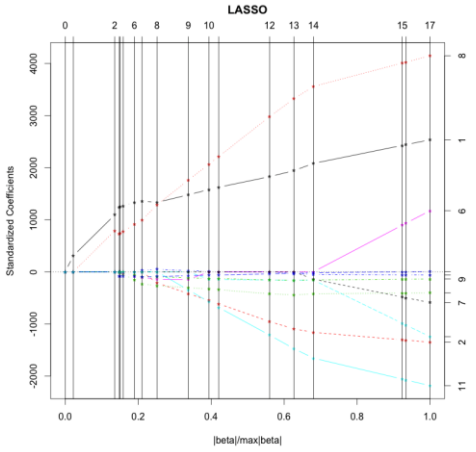
The important coefficients found by each method for NY are.

Rank	Stepwise	LASSO
1	LTV	Cap Rate
2	DSCR	Occupancy
3	Prop Type	LTV
4	Loan Purpose	Loan Purpose
5	Occupancy	Debt Yield

LTV and DSCR are financial measures directly linked with the owner's ability to repay the debt. As expected, their coefficients are positive. Occupancy determines the owner's ability to generate income; so does property type.

5.1.2 California

For California, besides these directly linked features, originator is included in the list of important features. Here is the LASSO path:



The following features were selected as top ranked by the methods.

Rank	Stepwise	LASSO
1	LTV	LTV
2	DSCR	Debt Yield
3	Prop Type	NOI
4	Originator	Prop Type
5	Loan Purpose	Loan Purpose

5.2 SVMs and Trees

We first present the results from SVM regression for both New York and California. The parameters were roughly chosen to minimize the cross-validated MSE.

The following table contains the cross-validated MSEs for each type of kernel.

	New York	California
Radial	3623.908	2881.713
Linear	4173.748	2590.822
Polynomial (deg. 2)	3427.33	2328.741
Polynomial (deg. 3)	3676.192	2478.713

It is clear that in both data sets, polynomial kernels, in particular the polynomial kernel of degree 2, works best. Hence, it is likely that there are interactions effects in the data.

With that in mind, we next present, for both New York and California, regression trees pruned to have the lowest cross-validated MSE and results from random forests.

5.2.1 New York

The following is the regression tree for the New York data set[4].

- root (250.13)
  - Originator=BANA, Beech Street Capital, CBRE, Centerline, CGMRC/GACC, CIBC, HSBC, JPM, Morgan Stanley, NCB, North Marq. Capital, Walker & Dunlop, Wells Fargo (213.26)
  - Originator=AMF, Barclays, Basis Real Estate, CCRE, CIIICM, Citigroup, GACC, Goldman Sachs, JLC, Key, LCF, NREC, RAIT, RBS, Redwood, UBS (284.84)
    - Property type[5]=MH, MU, OF, RT, SS (266.33)
    - Property type=IN, LO, MF, OT (329.26)
      - County=Bergen, Hudson, Nassau, New York, Ocean, Richmond (261.833)
      - County=Bronx, Essex, Kings, Morris, Queens, Suffolk, Union (379.83)

The numbers in parentheses are the average spreads at those nodes. This result is somewhat surprising; Lodging, the riskiest property type, and Multifamily, the safest, are in the same node. In addition, it seems that there is great variation in spread between counties, with more affluent counties having lower average spread.

The MSE of this tree was 4163.212. Random forests did improve this. Using 500 trees, the MSE was 2917.563.

The following table presents the 5 most important variables obtained by regression trees and random forests.

Rank	Regression Trees	Random Forests
1	Originator	Originator
2	Property type	DSCR
3	DSCR	Property type
4	LTV	LTV
5	County	County

Therefore, originator, property type, and DSCR appear to be the three most important variables for New York.

5.2.2 California

The pruned regression tree for California was simply the root node, with an average spread of 237.3 and an MSE of 6400. The random forests

[4] We give it in line form since the lettering on the graph is too small.

[5] MH: Manufactured Housing MU: Mixed Use OF: Office RT: Retail SS: Self Storage IN: Industrial LO: Lodging MF: Multifamily OT: Other

method considerably improves performance, with an MSE of 3680.926.

The following tree contains the 5 most important variables from the unpruned regression tree and random forests.

Rank	Regression Trees	Random Forests
1	Property type	Originator
2	Debt Yield	Property type
3	Occupancy	Occupancy
4	Cap rate	DSCR
5	Originator	Debt Yield

These rankings are much less regular than those in New York. However, it appears that overall, property type, originator, occupancy, and debt yield are important.

5.3 Aggregated Defaults

Reviewing the ex-post default statistics with the risk assessment implied in the average spread per property type gives a gauge of how linked average spread is to observed risk.

Rank	Avg. Spread	90 days delinquent	Fore-closures
1	Lodging	Multi Family	Multi Family
2	Industrial	Other	Industrial
3	Multi Use	Retail	Office
4	Multi Family	Industrial	Retail
5	Retail	Office	Lodging

We observe that Lodging was on average considered riskiest at loan origination, as expressed in the average spread of the loans, however, the category is surpassed by other categories in the ex-post assessment. We expect the evaluation criteria of new loans and with that the ranking spread of new loans will adapt to these outcomes.

6 Conclusion

One interesting observation is that the originator figures among the influential features under several of the methods. This implies that business practices of individual originators have considerable influence on loan terms. However, traditionally it is not considered in risk assessment.

As expected, the property type was also one of the influential features in the majority of our analysis. Yet, judging from the default data, the viewpoint that Lodging is the riskiest may be incorrect.

In both states, LASSO/Stepwise give differing results for variable importance than regression trees and random forests. In New York, both of the traditional metrics used in risk assessment, LTV and DSCR, were returned as important features. However, in California, only LTV was. This suggests that the market in California is unusual, and that banks should take other features when assessing the risk of loans there.

Another finding to support this is that random forests worked best for New York, but Stepwise worked best for California. This indicates that there are interaction effects in New York, but not in California.

7 Acknowledgements

We would like to thank Keith Siilats, who suggested this topic, for obtaining the data, answering our questions, and giving us guidance throughout the project.

References

[1] Trepp, LLC. *Treppv9* (2013) [Data file]. https://www.trepp.com/

[2] Hastie, T., Tibshirani, R, and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

[3] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*. New York: Wadsworth.

[4] Tibshirani, R. The LASSO Page. Retrieved December 1, 2013 from http://statweb.stanford.edu/~tibs/lasso.html.