Predicting Property Loan Spread Using Segmented Linear Model

Final Report

Sheng Zou (05804387), Jack Huang (05793331) and Yu Zhu (05731596)

***Abstract*—This project attempts to predict the interest spread of a property loan based on the borrower and property related attributes. Each attribute can be regarded as a potential feature. The problem is how to predict the spread accurately based on those features. This report describes our approaches of using linear and segmented linear models as well as other clustering methods. The comparative results and some analysis are also provided.**

Keywords—Loan Spread; LTV; DSCR; Segmented Linear Regression; MARS; K-means Clustering; PCA

I. INTRODUCTION

In debt and mortgage market, there are a number of pricing schemes available for determining the loan interest rate. In particular, many loan originators use risk-based pricing methods. Risks are generally composed of systematic risk, which is the economy-wide risk, and unsystematic risk, which is the company and investment specific risk. In this project, we attempt to design a model that can effectively capture how the unsystematic portion of the risk in a loan is related to some key attributes of the borrower and the property itself.

The loan originators tend to be risk averse, especially post the global financial crisis, and it is expected that for higher risk classes, the penalty on the interest rate should be higher than the lower risk classes. We use the loan spread as the measure of unsystematic risk. The loan spread is the difference between loan rate and treasury rate. Therefore, by subtracting treasury rate out we can eliminate the macro-economic factors that affect the nominal loan interest rate, which is the systematic risk.

Since we do not use any specific non-linear economic model as our prior knowledge, it seems natural to fit those attributes via some simple and extended linear models.

II. DATASETS

Our dataset consists of the fixed-rate property loan records in New York from Oct 2012 to Jun 2013, with benchmark at treasury-5-year, treasury-7-year and treasury-10-year rate. Table I provides a snapshot of a subset of the records contained in the dataset, and only some attributes of the records are shown. All the property records are obtained from the U.S. Security and Exchange Commission website.

As seen in Table I, the dataset has both numerical and categorical attributes.

TABLE I. A SNAPSHOT OF NEW YORK PROEPRTY LOAN RECORDS

Spread	DSCR	LTV	Year Built	Debt Yield	Property Type
292	1.41	62.3	2005	9.4	RT-Single Tenant
207	2.26	55.6	1892	10.1	RT-Unanchored
88.5	7.59	10.87	1957	46.82	CH (all)
255.5	1.28	65.22	1927	7.94	MU (all)
197	2.21	52.63	1927	9.07	OF-Urban
270.5	1.41	59.8	1917	8.8	SS (all)
251.5	1.79	70.48	2008	10.87	LO-Full Service

III. FEATURES

Historically, the most important fields in pricing the loan spread are LTV (loan-to-value ratio) and DSCR (debt service coverage ratio). LTV is the ratio of a loan to the value of an asset purchased, representing the fraction of money borrowed. DSCR is the ratio of cash available for debt servicing to interest, principal and lease payments, denoting the ability of the borrower to pay his/her debt. In addition, we will also examine a broad range of numerical features that could potentially affect the loan spread, including borrower as well as property related features:

- Yield (rate of return) of the debt
- Current year expenses per unit
- Current year revenue per unit
- Remaining terms of the loan
- Allocated Balance
- Built year of the property
- Occupancy of the property.
- Number of units of the property
- Remaining terms of the loan.

In addition, we also use the qualitative features to categorize the model. One such attribute is the property type, which potentially influences loan originators' perspective on the risk profile of investing on such property. For example, hotels are generally more risky than multi-family housing, as the returns of hotels are more difficult to predict. We can thus

divide the training examples into different categories and derive a separate regression model on each category.

IV. Data Preprosessing

To account for the difference in benchmark, we need to adjust the loan spreads by subtracting 110bps from the treasury-5-year loans and 60bps from the treasury-7-year loans, as the average gap of treasury rate between treasury-5-year and treasury-10-year are around 110bps from Oct 2012 to Jun 2013, and 60bps between treasury-7-year and treasury-10-year. This procedure further eliminates the systematic risk.

Before training the model, we selected the training samples with LTV between 10 to 80 and DSCR between 0 to 5 as the data points with values outside the above range are extremely rare and considered as abnormal and thus should be excluded from the pricing analysis.

V. Models

A. *Linear Regression: Spread vs. LTV and Spread vs. DSCR*

As our first attempt, we have constructed linear models for loan spread versus LTV and DSCR respectively. We use RMSE (normalized root-mean-square-error) to evaluate the effectiveness of the regression model. The RMS error is normalized by the mean of the spread, a constant for all models, which gives us a sense of how big the error is. Fig. 1 shows the 1-D linear regression of loan spread vs. LTV. The positive correlation observed is reasonable, as the higher the LTV value, the higher the risk and thus the higher the loan spread.

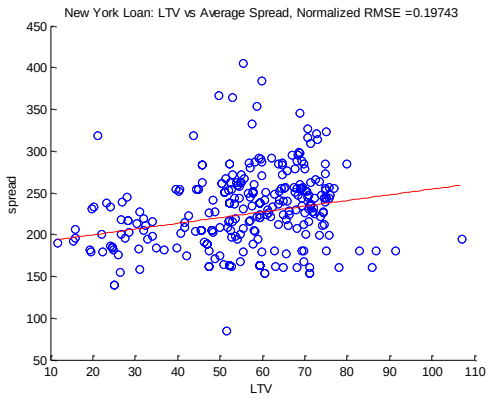

Fig. 1. Linear Regression of Spread vs. LTV

Fig. 2 shows the 1-D linear regression of loan spread vs. DSCR value. The negative correlation observed is reasonable, as the higher the DSCR value, the lower the risk and thus the lower the spread.

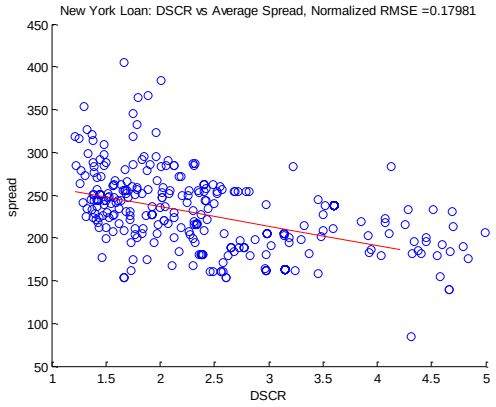

Fig. 2. Linear Regression of Spread vs. DSCR

B. *2-D Linear Regression: Spread vs. LTV & DSCR*

After observing the correlations of the two features separately, we then combined the two features to construct a 2-D linear model. It is observed in Fig. 3 that the normalized RMSE value only improved slightly from the 1-D model.

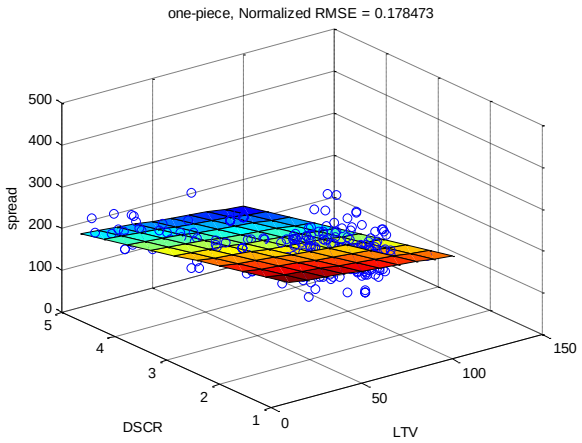

Fig. 3. Linear Regression of Spread vs. (LTV, DSCR)

C. *2-D Segmented Linear Regression*

We attempt to improve our regression model by segmentation. For the purpose of prediction, the model needs to be continuous at the boundaries. The breakpoints are manually set at midpoint: LTV = 40 and DSCR = 2.5 to divide the data into four regions. CVX is used for minimizing the least square error given the above constraints. As shown in Fig. 4, the RMSE value further improved from the previous 2-D linear model.

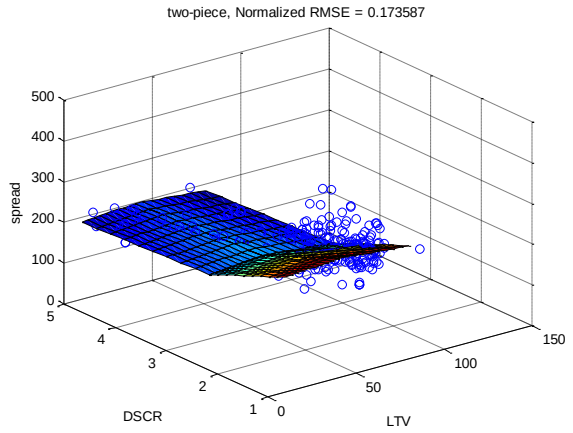

Fig. 4. 2x2-Piecewise 2D Linear Regression of Spread vs. (LTV, DSCR)

D. Multivaraite Adaptive Regression Spline (MARS)

In order to do more segments, we look at the multivariate adaptive regression splines (MARS) algorithm [1][2]. MARS has the ability to find the optimal placement of breakpoints using heuristics and perform piecewise linear regression. It also implements regularization to help avoid overfitting. Fig. 5 shows the result of using MARS, which further improves from linear regression and 2-by-2 segmented regression model. The equation for the model is given as:

$$y = -234 - 193\max(0, 3.6 - x2) + 557\max(0, 3.15 - x2) - 1.12\max(0, x1 - 58) - 341\max(0, x2 - 3.14) - 82.4\max(0,2.02 - \mathrm{x2}) + 290\max(0, \mathrm{x2} - 1.44)$$

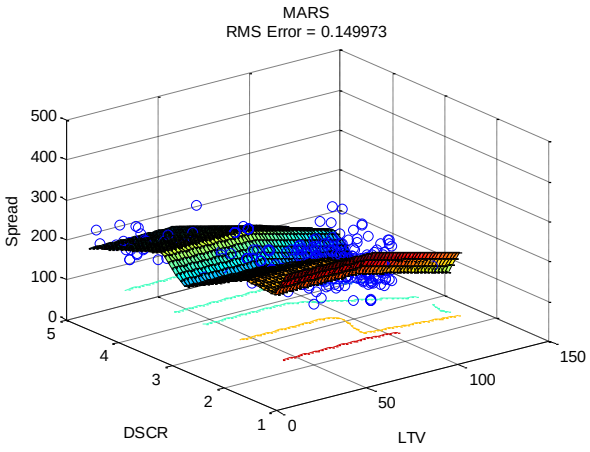

Fig. 5. MARS of Spread vs. (LTV, DSCR)

E. K-Means Clustering

A cluster of data with similar feature values should have similar risk profile. Therefore, we also attempt to cluster the data using k-means clustering algorithm and do a linear regression for each cluster. However, due to the limiting size of our dataset, which has 365 records, the smallest cluster for k=4 has only less than 40 data points, which potentially results in higher variance. The clustered data points and centroids according to LTV and DSCR values are shown in Fig. 6. It can be seen that the RMSE do not improve from using MARS.

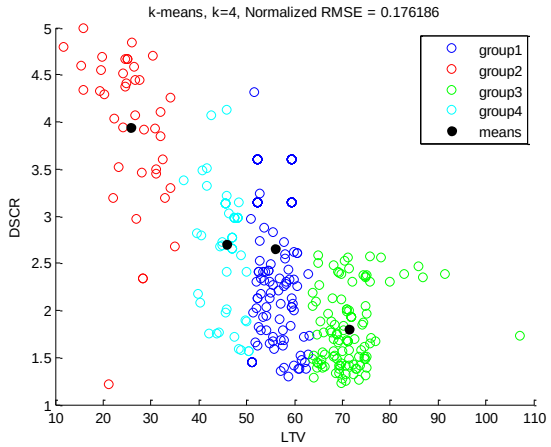

Fig. 6. (LTV, DSCR) Clustered Using K-Mean Clustering (k=4)

F. Property Type Clustering

We suspect that the risk profile for different property types will be different. Therefore, it comes naturally to generate different regression models for different property types. We first pick out some major types, which has a record size greater than 30 and then categorize the loans according to the types. A category "other" is created for those records which belong to minor types. This prevents us from having a category with too few records. Fig. 7 shows the categorized data points with their LTV-DSCR values (CH – Coop Housing, LO – Lodging (hotels), MF – Multi-family Housing, OF – Office, RT – Retails).

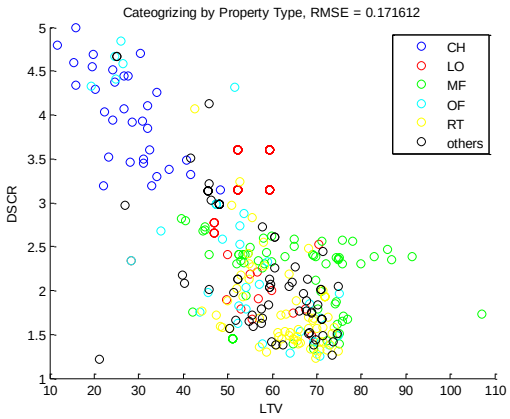

Fig. 7. (LTV, DSCR) Clustered Using Property Type

Fig. 8 compares the linear regression models for LO and MF respectively, it is clearly shown that for similar values of LTV and DSCR, LO loans generally have higher spread. The two models are therefore very distinct from one another.

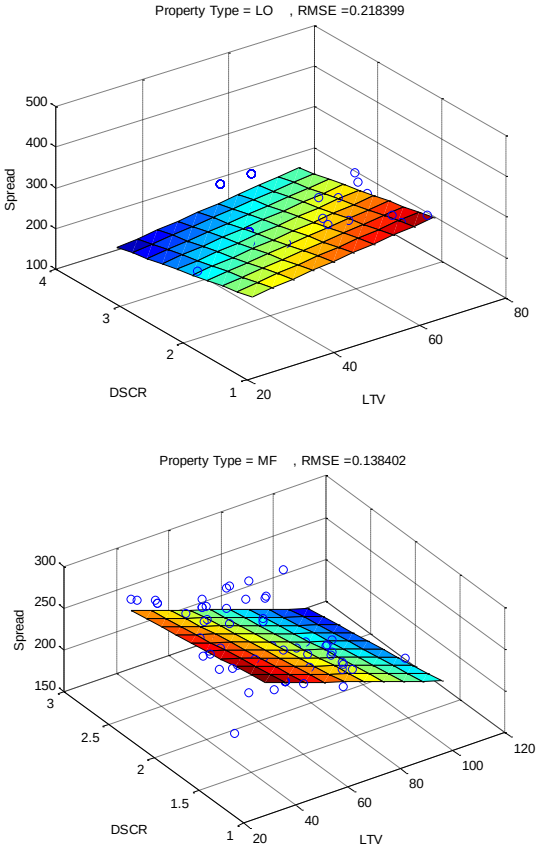

Fig. 8. LO Property Type Regression Model (Top) and MF Property Type Regression Model (Bottom)

G. Principal Component Analysis (PCA)

So far only DSCR and LTV are considered as the input numerical features because of their conceived strong correlation with the spread. However, the data file has many more numerical features, such as "Year Built" and "Debt Yield", as shown in Table I. We would like to use these features as well, and doing a PCA is a natural way to visualize the result.

First, we examined the data file and picked out nine other numerical features which might be useful. This prevents us from including some obviously improper features, such as "ZIP Code" (While spread might be correlated with geographical location, it is unlikely that the relationship between spread and ZIP Code is linear). Now, the standard PCA procedure is performed: The mean value is removed from each feature, and each feature is then normalized. The features are compressed down to two dimensions, and used for linear regression.

Fig. 9 shows the linear regression result.

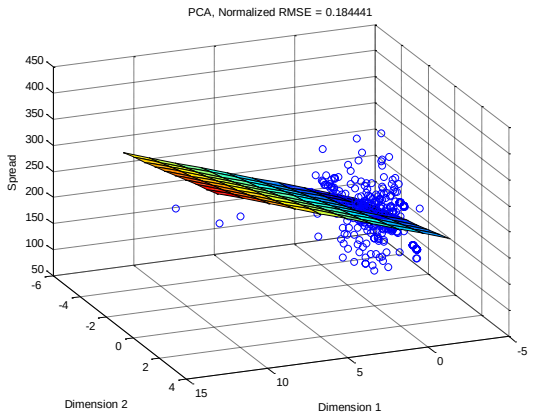

Fig. 9. PCA Linear Regression of Spread vs. (LTV, DSCR)

H. Mixture of Models

Finally, we combine the method of k-mean clustering, property type clustering, and PCA with MARS to test their performance. These more complex models can reduce the training RMSE value significantly, as shown in section VI. However, due to the substantial increase in fitting parameter numbers, we expect the variance of our model to be very high, which is illustrated in the large gap between training error and generalization error in cross-validation analysis described in the next section.

VI. Results and Analysis

We used ten-fold cross validation to evaluate each of our models, for the cases of two input features (LTV, DSCR) and eleven input features (including other nine numerical features). The results are shown in Table II.

TABLE II. CV Results for Different MOdels

Normalized RMSE	**Two Input Features (LTV, DSCR)**		**Eleven Input Features**	
	Training Error	Prediction Error	Training Error	Prediction Error
Mean (Baseline)	20.2%	20.1%	20.2%	20.1%
2D Linear	17.8%	17.9%	16.3%	17.1%
2x2 Linear	17.2%	17.4%	9.1%	83.5%
Mars	15.0%	15.4%	12.7%	113.4%
PCA	NA	NA	18.8%	20.4%
K-means	17.5%	18.1%	12.9%	78.7%
Type	17.4%	26.3%	12.7%	125.1%
PCA-Mars	NA	NA	15.1%	22.5%
K-means & Mars	13.7%	16.3%	9.8%	78.7%
Type & Mars	12.4%	17.9%	9.1%	83.5%

The first row, "Mean", means we just predict the spread to be the average of the spread in the training data and use it as the baseline to evaluate the performance of other models.

For two input features, we see that the training error is generally decreased by using more complex models. In contrast, the gap between training error and generalization error are

wider when more parameters are introduced, indicating higher variance and the training data is overfitted. This is reasonable because those models contain higher VC dimensions,

When we increase the input features to eleven, the problem of overfitting becomes more obvious, as all models except Linear suffer from huge variance. One exception is PCA and PCA & MARS, since PCA reduces the number of features back down to 2, its variance is significantly smaller than the rest.

However, an interesting observation is that the generalization error of using PCA is still worse than simply using LTV and DSCR directly for linear regression. This implies that PCA should be used for a good reason and it might not outperform the result of simply picking the best features.

The best performing model is MARS with 2 input features. This hints that a linear model cannot capture very well the relationship between the spread and the features, and there might be some deep nonlinearity inside which is beyond the goal of our project.

Compared with using MARS alone, the inferior performance of K-means and Type in both training and generalization error is beyond our expectation. However, considering the result given by MARS, this is not impossible. Because of the inherent nonlinearity, putting our data into different types does not help overcome this problem. On the other hand, clustering and categorizing might lead to highly uneven splitting of the data and some groups might have very few points. This will cause potential high variance. Generally speaking there should be a tradeoff between bias and variance, however, in the case of strong nonlinearity and if our model does not take that into account, both bias and variance can be high.

Another thing that is worth mentioning is that our data is very noisy, despite the nonlinear structure it has. This is because the loan originator's decision might also be influenced by subjective factors, and cannot be simply predicted by limited open objective observation. Such large noise will tend to make our models have large variance given few data.

Finally, the loan records have the following characteristic: Many records have missing fields. This implies a large portion of the data is given only incomplete set of features. Our approach is to substitute the missing fields by the average value of those records in the training set which have valid feature values. However, the actual unobserved fields might be far from the given average.

VII. CONCLUSION

Give a data file of loan records, we tried to predict the loan spread, which is a measure of the unsystematic risk, by using some attributes from the borrower and the property. We tried various methods and found that using an adaptive segmented linear model called MARS with two most important features, i.e., LTV and DSCR can give the best performance. This is because the adaptive segmented linear model can partially describe the non-linear structure in the data, and using only two features minimizes the risk of overfitting.

Clustering and categorizing are still potentially promising, and in high dimensional case combined with MARS the training error is dramatically reduced. However, because of increased variance, the ultimate performance is not so satisfying. It still needs to be examined whether increasing data points can improve their generalization error.

The large noise observed in the loan spread is an indication that the spread is not well modelled by the given features. Looking for more relevant features is a crucial task to address this problem.

The last challenge is how to manipulate partially observed data in our training model. There might be better method than replacing missing fields with observed mean values.

VIII. ACKNOWLEDGEMENT

We would like to thank Keith Siilats for his help in getting us started, offering data, and providing suggestions along the way.

IX. REFERENCE

[1] Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines," The Annals of Statistics 19: 1.

[2] "Open source regression software for Matlab/Octave," URL: http://www.cs.rtu.lv/jekabsons/regression.html

[3] Andrew Ng, "CS 229 Course Notes," URL: http://cs229.stanford.edu/materials.html