Improving Lending Through Modeling Defaults

BUDT 733: Data Mining for Business
May 10, 2010

Team 1
Lindsey Cohen
Ross Dodd
Wells Person
Amy Rzepka

EXECUTIVE SUMMARY

Background

Prosper.com is an online peer-to-peer lending system for borrowing money and investing in loans through an open and transparent auction model. Prosper.com borrowers create credit profiles containing information lenders can review before determining whether to invest or not in a borrower. Even with this information, one challenge Prosper.com lenders face is being able to predict which borrowers will default on loans.

Goal

The goal of this project is to assist with lending decisions by creating a model for Prosper.com lenders that will classify new listings as according to whether or not they are likely to default.

Data

To accomplish our goal we downloaded publicly available data from Prosper.com. The data gathered is a complete snapshot of all listings created on Prosper from November 2005 to January 2010. The data includes information on listings, loans (listings which have become loans), group membership within Prosper.com and cross-referenced categories. After merging the files and filtering for completed loans there were 19,509 loan records and 39 predictors (categorical and numerical).

After several exploratory studies and with the help of domain knowledge, the final predictor list was narrowed to the following predictors: Amount Requested, Borrower Max Rate, Borrower State, Listing Category, Credit Grade, Debt to Income Ratio, Description, Duration, Funding Option and Group Rating.

Model Selection

Models were developed using the following methods: Logistic Regression, K-Nearest Neighbors (KNN) and Classification Tree. We decided on a logistic regression model with 22 variables based on 10 predictors because it had close to the lowest default error rate (2.37%) and overall error rate (38.35%) for the test data. The KNN model also had a very low default error rate, but was more complicated than the logistic regression model and had a larger overall error rate.

Recommendation

For lenders looking for an in-depth and accurate model we recommend the logistic regression model. A lender can utilize the information produced from this model to create a subset of potential loan listings to bid on. However, it is important to note that while the classification tree's default error performance was not ranked at the top it did have the best no default error rate. Also, for lenders looking for a simple and transparent model the classification tree is a viable option. However, in the end the final decision on which of these loans to bid on is left to the discretion of the lender.

TECHNICAL SUMMARY

Goal

The goal of this project is to create a model for Prosper.com lenders that classifies listing based on whether or not they are likely to default in order to assist with the lending decision process.

Data Preparation

We turned the "Status" variable into a binary response variable with "Default" and "No Default" as our 2 classes. However "Status" originally had 14 categories, so we had to determine how to bin the different statuses into either "Default" or "No Default." Using our domain knowledge and research from Prosper.com we determined the following classification: Default = any loan that was late, defaulted, repurchased or charged-off; No Default = payoff in progress and paid. By classifying the statuses in this way we were conservatively classifying records and erring on the side of caution, which we felt was reasonable. Current and cancelled loans were omitted because we could not yet evaluate whether or not they have or would have defaulted, thus they could not be used in our model. After removing those records we ended up with 19,509 observations.

Our initial cleanup consisted of deleting duplicate columns that were a result of the merging. We then deleted those predictors which would not be known at the start of the bidding process or had no meaning (i.e. unique ID keys). Next we searched for erroneous and missing values. We found two observations where typos were apparent and fixed them. Our remaining search resulted in five predictor columns that contained records with missing values. We used various methods to deal with these missing values. We chose to delete one of the predictors because we felt the missing information was captured in another variable. Thus this predictor added no additional value and it could be deleted. Upon further investigation we found that data for one of the predictors was not recorded until 2009 so we chose to delete that variable as well. Next we turned one of the predictors into a response/no response variable because we felt the missing responses may offer some insight. Lastly, we used our domain knowledge to impute missing values for two of the variables.

Data Exploration

We spent a significant amount of time exploring the remaining set of variables looking for relevant predictors. We used a combination of Spotfire and Excel for data exploration and visualization. A series of box plots, scatter plots and pivot tables were generated to explore the data. Some of the charts explored are shown in Appendix A. Those variables which did not exhibit any separation were eliminated. Our initial exploration revealed 11 predictors of significance. They are as follows: Amount Requested, Draft Fee, Borrower Max Rate, Borrower State, Listing Category, Credit Grade, Debt to Income Ratio, Description, Duration, Funding Option and Group Rating. Please see Appendix B for variable definitions. Since many of our variables were categorical we converted them into dummies. However this resulted in a large number of variables. So to further reduce this number we used pivot tables and Spotfire charts to look for classes within categories that had similar distributions. We then determined the appropriate number of bins for each category and in doing this we reduced our number of variables to 34.

Model Creation & Selection

We first partitioned the data into training (50%), validation (30%), and test (20%) datasets because some of the models we used (classification tree and KNN) used the validation set to optimize the initial model. Four different models were considered: Discriminant Analysis (DA), Classification Tree, KNN and Logistic Regression however only three of the methods were run. We rejected DA

as a viable method for our predictions due to our numerical variable not being normally distributed (1 of the 2 assumptions that must be met to use DA).

For all of our models we initially ran them with a cutoff of 0.5 and "Default" as the success class. However, since our goal is to find and properly classify loans that will default, we reduced the cut-off from 0.5 to 0.2 in all our models. In doing so we were able to drastically improve the error rate for classifying a loan as "Default." Given that in the test data there is over $2.5M in loans predicted to not default at the 0.2 cutoff level, we still believe overall there is a sufficient amount of listings available to be invested in.

K-Nearest Neighbor (KNN)
We ran the KNN model using all 34 variables. Using the test data we arrived at a default error rate of approximately 2.27% and an overall error rate of 41.52%. Please see Appendix C for the results.

Classification Tree
We ran the classification tree using all 34 variables, however the best prune tree used "Borrower Max Rate" and "Amount Requested" as the predictors. Looking at the test data using the best pruned tree the default error rate was 7.16% and the overall error rate was 37.67%. Please see Appendix C for the results.

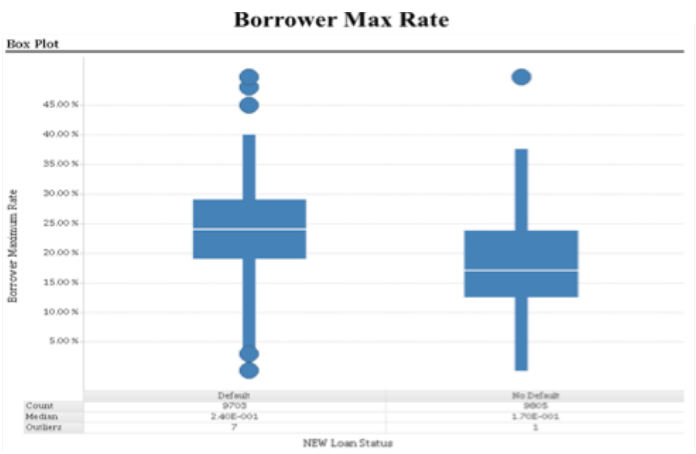
Logistic Regression (LR)
We first ran a logistic regression with 11 predictors and 34 input variables. Using stepwise regression in XL Miner we examined the best subsets. We chose a model with 22 variables because the Cp value was close to the number of variables and there was a fairly big jump in RSS value. The error rate for both models were similar, so in the interest of parsimony we felt the model with few variables was best. Using the test data from this model we arrived at a default error rate of 2.37% and an overall error rate of 38.35%. Please see Appendix C for the results.

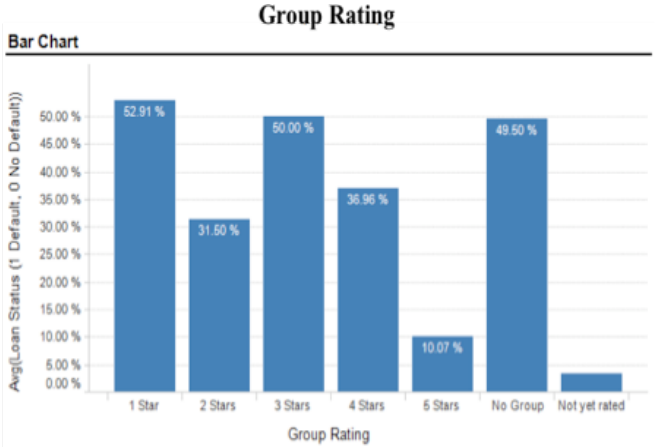
Model Selection
We decided on a logistic regression model with 22 variables based on 10 predictors because it had close to the lowest default error rate (2.37%) and overall error rate (38.35%) for the test data. The KNN model also had a very low default error rate, but was more complicated than the logistic regression model and had a larger overall error rate.

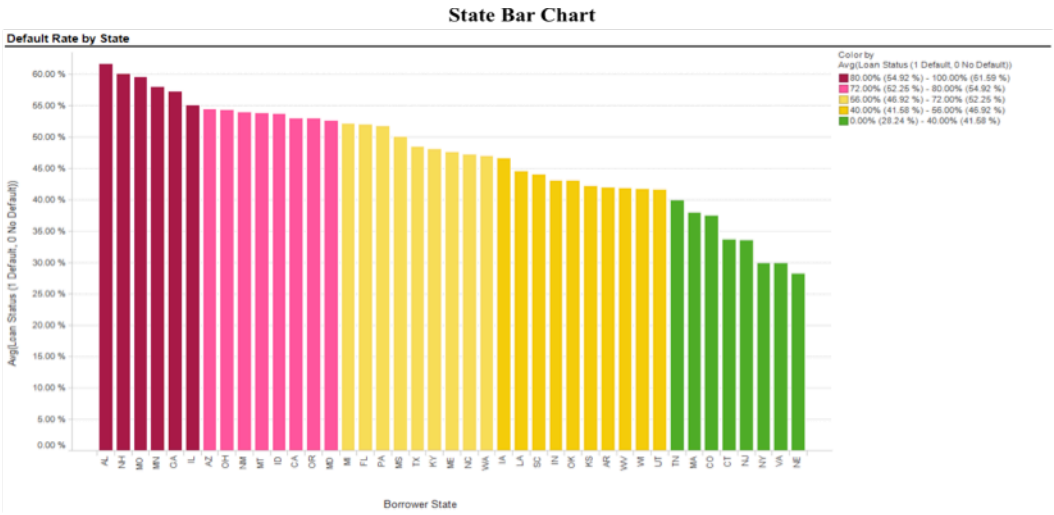
Recommendation
While the classification tree did not yield the best error result rate for predicting those who will default it did yield the best error result rate (67.91%) for predicting those lenders who will not default as well as the best overall error rate (37.67%). Therefore this model is still a viable option. The tree is also useful to those lenders who are looking for a relatively simple, "off-the-shelf" predictor. The tree has a practical advantage in that it uses few variables and helps generate a transparent set of rules. Thus for those lenders considering a large number of candidates they could quickly classify those candidates using the tree.

For lenders looking for a more in-depth and accurate model we recommend they use the logistic regression model to create a subset of potential loan listings to bid on. While the final decision on which of these loans to make a bid on is left to the lenders discretion these models should aid in increasing the lender's return.

APPENDIX A – Data Exploration







FUNDING OPTION		
Count of Loan Status	NEW Loan Status	
FundingOption	Default	No Default
Close When Funded	58.30%	41.70%
Open For Duration	45.84%	54.16%
Grand Total	49.74%	50.26%

APPENDIX B – Variables

VARIABLE	DEFINITION
Amt Requested	Amt the member requested to borrow
Bank Draft Fee	Indicates whether or not a bank draft fee was charged
Borrower Max Rate	Max interest rate borrower is willing to pay
Borrower State Risk	Bins borrowers into groups based on their state
Listing Category	Category of the listing
Credit Grade	Bins borrowers into categories based on their credit grade
Debt to Income Ratio	Bins borrower's into groups based on their debt to income ratio
Description	Specifies words contained in the listing description
Duration	Bins borrowers into groups based on # of days listing is valid for
Funding Option	Indicates whether or not the listing closes after it has been funded
Group Rating	Bins borrowers into groups based on the group rating of a group

APPENDIX C - Models

Logistic Regression

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.95106745	0.2474373	0	*
Amount Requested	0.00007247	0.00000458	0	1.00007248
Borrower Max Rate	7.25551128	0.39828181	0	1415.886841
Borrower State Default Risk_H	0.10550274	0.05537205	0.05673551	1.11126912
Borrower State Default Risk_L	-0.23977007	0.07621536	0.00165546	0.78680873
Listing Category_1	0.57923812	0.0771006	0	1.7846781
Listing Category_2	0.58838588	0.18368024	0.00135855	1.80107892
Listing Category_3	0.60873151	0.10972508	0.00000003	1.83809829
Listing Category_4	0.36570379	0.09747486	0.00017559	1.4415282
Credit Grade_Good	-0.76276892	0.07382566	0	0.46637329
Credit Grade_Poor	0.84771073	0.07155327	0	2.33429694
Debt To Income Ratio Cat_.1 - .2	-0.21709873	0.06386801	0.00067589	0.80485052
Debt To Income Ratio Cat_.2 - .3	-0.11284501	0.06822211	0.09811074	0.89328909
Debt To Income Ratio Cat_.5 - .7	0.29220024	0.12537257	0.01977155	1.3393712
Debt To Income Ratio Cat_.7 - 10.1	0.29083833	0.14132056	0.03958971	1.33754838
Debt To Income Ratio Cat_0 - .1	-0.32112059	0.07564927	0.00002187	0.72533578
Description Contains Sick, Illness , Cancer, Hospital, or Disease_YES	0.23090096	0.09096003	0.01113326	1.25973451
Description Contains Jesus, Christ, Angel, or God_YES	0.41159391	0.08862063	0.00000341	1.50922143
Duration_14	-0.92463613	0.25474563	0.00028381	0.39667574
FundOption_Close When Funded	0.37850854	0.05104887	0	1.4601053
Group Rating_Good	-0.60148919	0.29286799	0.03999608	0.54799497
Group Rating_Poor	0.83950984	0.22352892	0.00017285	2.3152318

Class	Prob.	
Default	0.493336067	<-- Success Class
No Default	0.506663933	

Residual df	9732
Residual Dev.	11360.4209
% Success in train data	49.33360673
# Iterations used	10
Multiple R-squared	0.1597435

Test Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Default	No Default
Default	1896	46
No Default	1450	509

Error Report			
Class	# Cases	# Errors	% Error
Default	1942	46	2.37
No Default	1959	1450	74.02
Overall	3901	1496	38.35

Classification Tree

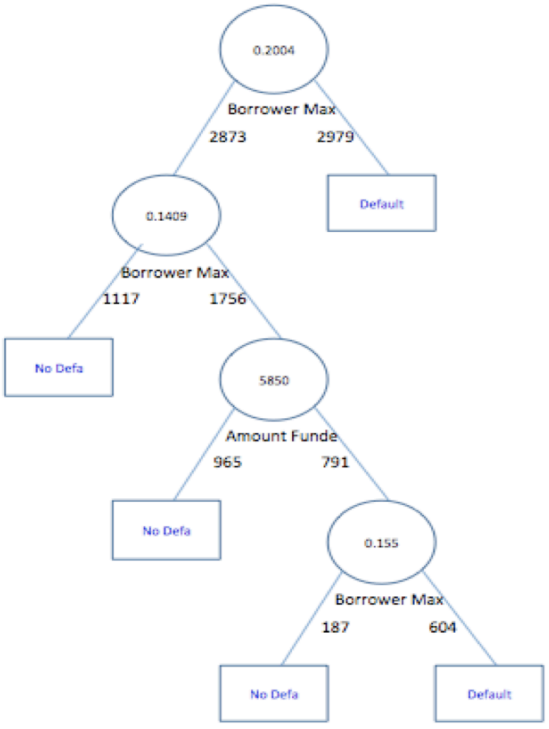
Test Data scoring - Summary Report (Using Best Pruned Tree)

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Default	No Default
Default	1803	139
No Default	1331	629

Error Report			
Class	# Cases	# Errors	% Error
Default	1942	139	7.16
No Default	1960	1331	67.91
Overall	3902	1470	37.67

Classification Tree - Best Pruned Tree (Using Validation Data)



K-Nearest Neighbors Classification

Test Data scoring - Summary Report (for k=20)

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Default	No Default
Default	1898	44
No Default	1576	384

Error Report			
Class	# Cases	# Errors	% Error
Default	1942	44	2.27
No Default	1960	1576	80.41
Overall	3902	1620	41.52