Predictive Model for Prosper.com

BIDM Final Project Report

Build a predictive model for investors to be able to classify "Success" loans vs "Probable Default" Loans

Sourabh Kukreja, Natasha Sood, Nikhil Goenka, Salil Das, Vikas Shah

12/23/2010

Contents

Executive Summary

Prosper is the world's largest peer-to-peer lending marketplace, with more than 1,020,000 members and over $213 million in funded loans. Borrowers' list loan requests between $2,000 and $25,000 and individual lenders invest as little as $25 in each loan listing they select. It rates all the borrowers' and using those ratings lenders decide if and how much they want to invest. Please refer to *Exhibit A* for more details.

Problem Description

We started the project trying to predict the prosper.com rating for any user that signed up to borrow money on the website. As we went through the data we realized that it would make a lot more business sense to build a model over the existing Prosper rating. In other words we wanted to see if we could improve investor's odds of investing in a possible defaulter and provide tools for better decision making by using our model.

Traditional Way

At the moment when a user signs up as a borrower based on a series and historical data that is available on the user, prosper.com assigns them a credit grade. They started doing as recently as Nov '09. Till then they were using a default rate that the national credit rating agency assigns every individual in the US. This is a generic rating that all citizens are assigned. Prosper now uses their own system but there is still room for improvement especially because Prosper is very slow in assigning ratings to its users.

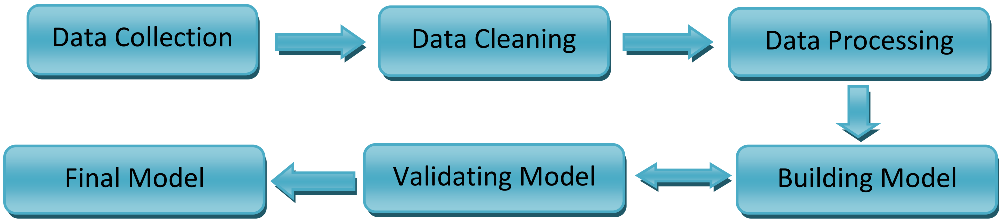
Proposed Way

The broad aim is to build a model that it will help investors predict whether a new borrower listing will result in an "on-time" payment or will it lead to a delayed payment or default. This predictive ability of the model will help investors decide whether they should bid for a certain listing or not and if they do decide to invest then what will be the chances of a default.

Currently, Prosper.com rating assigns "Prosper Ratings" to the borrowers to help investors make the same decision. However, there are many borrowers (about 50%) who have not been assigned these ratings and investors have no ratings to rely on for investing in these listings. This is where our model will be very useful – it can predict the rating for any borrower – existing or new. Moreover, even for borrowers who have been assigned Proper Ratings, if investors use our model along with the given ratings, the accuracy of the predictions will improve.

Thus, the model will provide a way to predict whether a new loan will default or not with a high degree of accuracy.

Process Followed



Data Management

We collected over 1.5 GB of Data from prosper.com containing over 2 million data rows in 5 different tables. The data collected was between 2006 and 2010.

Data Sources

We downloaded the data from Prosper.com export. The data export provides a daily snapshot of all the public data available in the Prosper Marketplace.

Data Schema

There are 5 main tables in which data about loans, lenders and borrowers are stored. Description of each of the tables and schema is described below.

Member Object: A Member is a registered user of the Prosper Marketplace site. A Member may have one or multiple roles that determine which actions the Member is allowed to perform on the site.

Listing Object: A Listing is created by a Borrower to solicit bids by describing themselves and the reason they are looking to borrow money. If the Listing receives enough bids by Lenders to reach the Amount Requested then after the Listing period ends it will become a Loan. A Borrower may only have one active listing at a particular moment in time.

Group Object: A Group is a collection of Members who share a common interest or affiliation. Groups are managed by Group Leaders who bring borrowers to Prosper, maintain the group's presence on the site, and collect and/or share Group Rewards. Borrowers who are members of a group often get better interest rates because Lenders tend to have more confidence in Borrowers that belong to trusted Groups.

Loan Object: A Loan is created when a Borrower has received enough Bids to meet the full amount of money that the Borrower requested in their Listing. The Borrower must then make payments on the Loan to keep it's status current.

Please refer to *Exhibit B* for Data Schema Diagram with Primary Key and Foreign Key constraints.

Data Processing

Following is brief description of the steps that we followed to preprocess the data we had before moving towards data analysis.

1) **Convert XML to CSV**: Given data was in Extended Markup Language (XML) and we had to use a tool to convert it to CSV to be able to proceed with our analysis.
2) **Merging Tables:** We needed to merge the tables to get one consolidated data table to work on. We had to take care of
   a. **Foreign Key Constraints:** While merging the tables we had to take care of foreign key constrains in the tables. The keys are highlighted in the schema described above.
   b. **Joins:** We also had to make tradeoffs between different type of joins while merging the tables. We decided to go with Left Outer Joins (Left being Listings table) to ensure that we do not end up with lots of missing values in other tables.
3) **Random Sample:** For our analysis we needed a random sample of about 50K rows from the 2 million rows of data. To achieve this we used a creative approach to adding one more column with random no generator and then picking up first 50K rows. This process ensured that our data sample was truly random
4) **Missing Data Values:** We used the median for the continuous missing data values and false for Boolean data values.
5) **Binning Data:** Data values like Status (with 10+ types of value) need to be combined and binned for our data analysis.

Data Visualization and Key Findings

Exhibit B shows some of the immediate findings that we got from the data set when we first visualized it. Some of the insights we got were

B1: 10% drop seen from Dec 2008 – Jan 2009 and Borrowers/Lenders ratio almost always 2:1 and we can see that lenders have power over borrowers.

B2: Loans closed rate grew significantly until Nov 2008 which coincides with the class action lawsuit. The curve follows the adopter's curve that is usually seen with new products.

B3: Having an endorsement increases your chance of being funded by 18%

Data Mining Models

The expectations from the data mining models are multifold:

- Firstly, highlight the factors/variables which play an important role in determining whether loan will be paid on time or not
- Secondly, understand how changes in these variables affect the outcome, i.e., understanding the sensitivity of predicted out come on the value of these critical variables
- Thirdly, build an algorithm where, by entering the values of these few critical variables, the outcome can be predicted

Since, most of our variables were categorical; we used logistic regression and classification tree to bring out the relationship between the key variable inputs and the output.

Logistic regression helped us understand which key variables does the predicted outcome depend on and how much variation in the predicted outcome can be explained by our model. It also helped us understand the accuracy of our model and the correlations between the different variables – i.e. how much noise or co-linearity is there between the different input variables.

Classification trees also helped us immediately visualize the relationship between the different input variables and the output. It showed us which are the most important variable that determine the output and how changes in those input variable will affect the output, i.e., sensitivity.

Logistic Regression Model

Predictors Used

- Bid Count
- Borrower rate
- Lender rate
- Age in months
- Amount Borrowed
- Is Home Owner
- Debt To Income Ratio

Output

"Status" with two categorical classes: Default /Late or Paid

Model Output

Based on the regression output, we were able to narrow down the list of important variables to about 6 predictors: BorrowerRate, IsBorrowerHomeowner, LenderRate, AgeInMonths, AmountBorrowed, and Term (refer Exhibit D).

Co-relation Analysis

All the predictors are uncorrelated expect the BorrowerRate and LenderRate which was expected as they are always interlinked. Thus, we were able to come up with a bunch of uncorrelated variables to explain the variation in the output (refer Exhibit D).

Goodness of Fit: Performance & Error Rate

The model performed very well in predicting which loans will be paid on time with an error of just 10% in both training and test data (refer *Exhibit D*).

Classification Tree

Predictors Used

- Bid Count
- Borrower rate
- Lender rate

- Age in months
- Amount Borrowed
- Is Home Owner
- Debt To Income Ratio

The parameters of the tree were: Best prune tree with 100 elements in terminal node.

Output

"Status" with two categorical classes: Default /Late or Paid

Model Output

The top three predictors that emerged from the classification tree were:

- Age in Months
- Borrower Rate
- Bid Count

Goodness of Fit: Performance & Error Rate

The model was able to successfully predict whether a loan will be paid on time with about 11% accuracy. Also, the model performs very well for higher Prosper ratings as can seen in *Exhibit E*.

Application of the Model

When we started the project we thought that out model should be able to predict a prosper rating for a listing. This would really help us create ratings for more than 50% of the records for which Proper ratings are missing. After the logistic regression and Classification tree analysis we identified real potential of model. Some of the observations and applications are

a) Credit rating and Prosper rating are mutually exclusive. Sometime in 2009 Prosper decided to get away with Credit Grades and come up with own Prosper rating system.
b) Not all (less than 50%) records after 2009 have prosper rating. This might be the reason why investors do have any guidance for making investments.

The real potential of our model was explored when we used our model along with the Prosper ratings.

a) <u>Model with prosper ratings:</u> Our model when used along with prosper rating has error rate of less than 2% to be able to identify "On-time" payment , and thus identify lemons from real investment opportunities.
b) <u>Model with Credit ratings:</u> We tested our model along with credit ratings and our predictive rate for identifying "successful" opportunities was 70% across AA – NC categories.

Please refer to the *Exhibit F* for Model accuracy with Prosper and Credit Ratings

Extensions possible for our model

Some of the proposed extensions of our model that we can try are

1) **Role of Social Network:** We highlighted earlier in our analysis that endorsements increase the chance of getting a loan by 60%; similarly we can find out the role network of friends play in getting a loan. This data is present in the Listing and Member tables but in complex HTML format. We can easily extend our model to analyze affect
2) **Group Categories:** We also can extend our model to take into account, Group Categories e.g. Religious / Ethnic etc. We expect that some categories such as Religious / Social Service to have greater chance of loan and "on-time" completion.
3) **Co-relation with State / Time:** We also believe that region / time of the year also plays an important role in loan approval and payments. We can easily extend our model to taken into account State and Listing creation information.

Conclusion

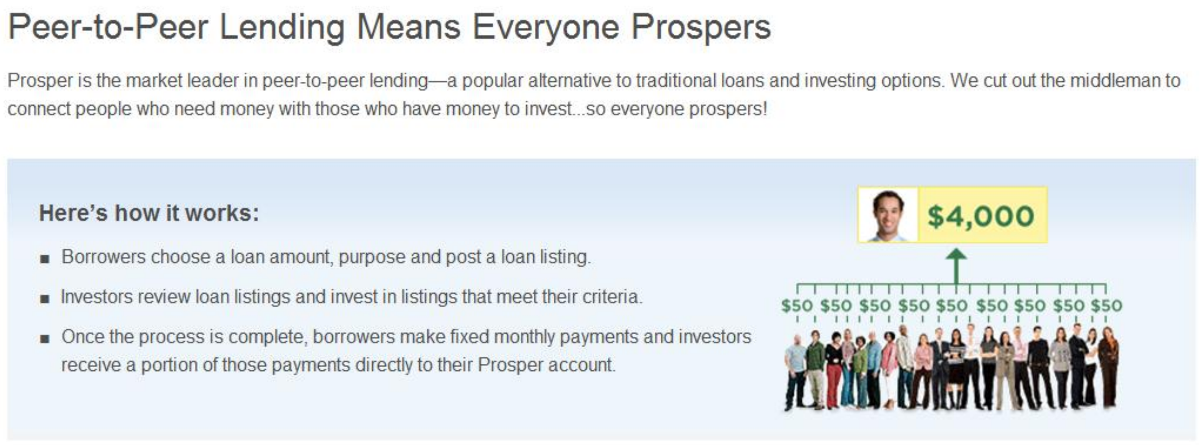
Our model when used with Prosper rating (if present) or in general any credit rating provided to the borrower provides a great tool to the investor to identify a good investment opportunity.

Monetization: With the extensions mentioned above, our model can really be monetized with small investors paying very small amount to get recommendations for investment opportunities on Prosper.com.

Exhibits

Exhibit A: Facts about Prosper.com from website / Wikipedia

Website:



Peer-to-Peer Lending Means Everyone Prospers

Prosper is the market leader in peer-to-peer lending—a popular alternative to traditional loans and investing options. We cut out the middleman to connect people who need money with those who have money to invest...so everyone prospers!

Here's how it works:

- Borrowers choose a loan amount, purpose and post a loan listing.
- Investors review loan listings and invest in listings that meet their criteria.
- Once the process is complete, borrowers make fixed monthly payments and investors receive a portion of those payments directly to their Prosper account.

Wikipedia:

Prosper Loans Marketplace, Inc. is a San Francisco, California-based company in the emerging peer-to-peer lending industry. The company operates Prosper.com, an online auction website where individuals can buy loans and request to borrow money. According to reports in the Wall Street Journal, "Prosper works like an eBay-style online auction marketplace, with lenders and borrowers ultimately determining loan rates."[1]

Prosper verifies selected borrowers' identity and personal data before funding loans[2] and manages loan repayment. These unsecured loans are fully amortized over three years, with no pre-payment penalty. Prosper generates revenue by collecting a one-time fee on funded loans from borrowers, and assessing an annual loan servicing fee to loan buyers. The idea for the service is derived from group banking concepts, such as rotating savings and credit associations. Other motivating ideas derive from the concept of microlending.

Prosper publishes performance statistics on the website; these are available to the public at large.[3] All transactions are in US dollars; lenders and borrowers must be US residents.

Prosper opened to the public on February 5, 2006. Prosper was founded by Chris Larsen, who also founded E-loan, and John Witchel and is backed by Accel Partners, Benchmark Capital, Fidelity Ventures, Omidyar Network, DAG Ventures, TomorrowVentures and Meritech Capital Partners.

Exhibit B1:

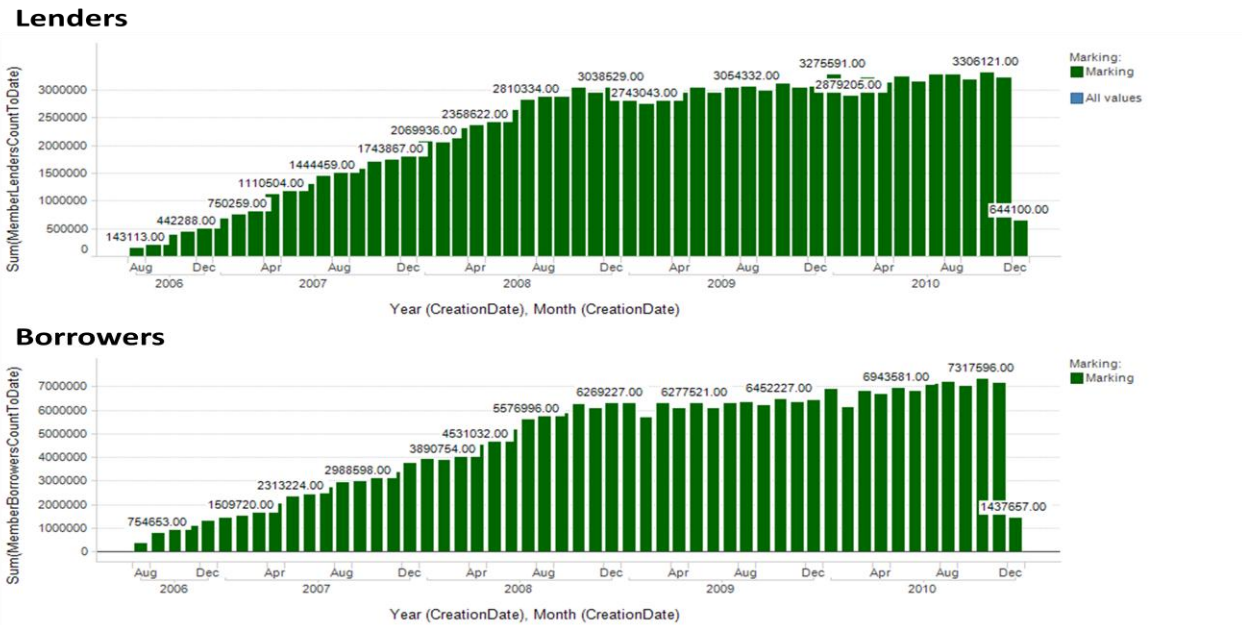


Exhibit B2:

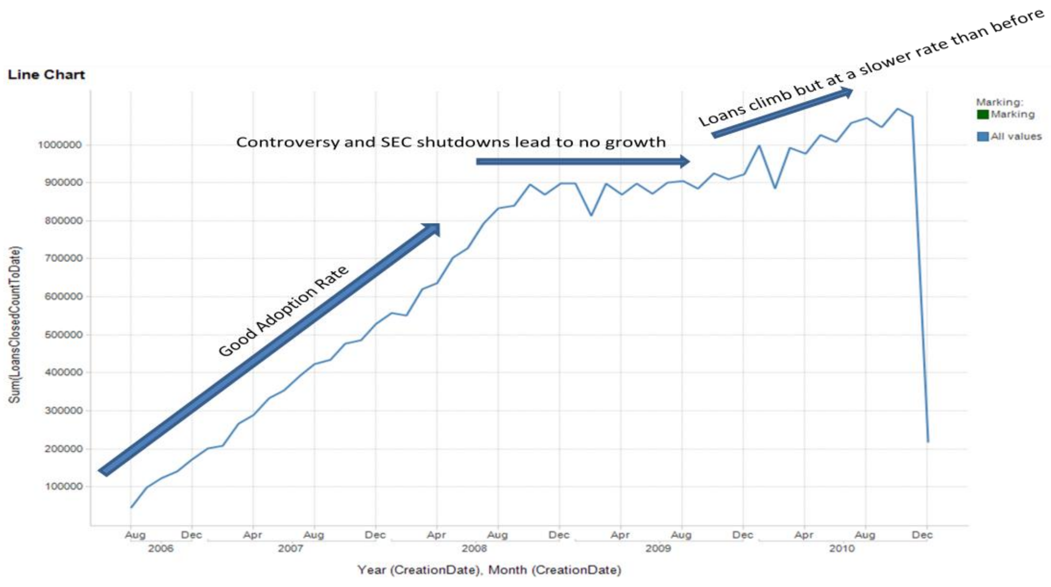


Exhibit B3:

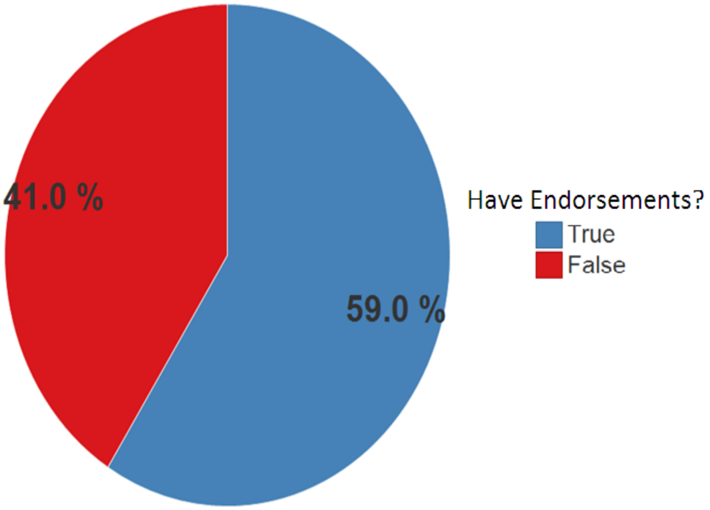
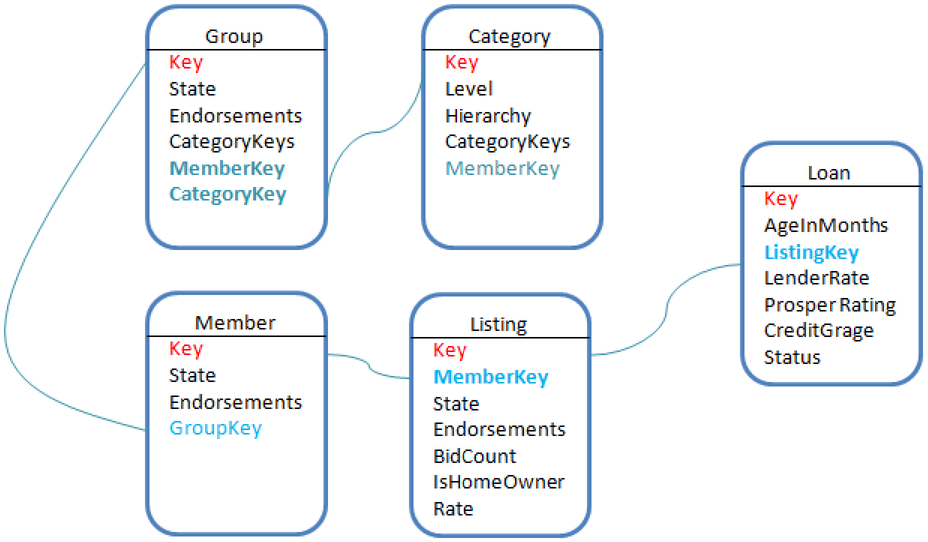


Exhibit C: Data Schema



Primary Keys highlighted in red and foreign keys highlighted in Blue.

Exhibit D: Output from Logistic Regression

Regression output:

Input variables	Coefficient	Std. Error	p-value	Odds
BidCount	0.00045379	0.00028042	0.10560962	1.000454
BorrowerRate	24.77111626	4.57036781	**0.00000006**	5.727E+10
DebtToIncomeRatio	0.03414476	0.02391896	0.15343051	1.0347344
IsBorrowerHomeowner	0.17194989	0.05088718	**0.00072739**	1.1876184
LenderRate	-15.33367443	4.61400652	**0.00088963**	2.2E-07
AgeInMonths	0.05707799	0.0022076	**0**	1.0587384
AmountBorrowed	0.00003511	0.0000068	**0.00000024**	1.0000352
Term	-0.141791	0.00396481	**0**	0.8678026

Correlation matrix:

	BidCount	BorrowerRate	DebtToIncomeRatio	IsBorrowerHomeowner	LenderRate	AgeInMonths	AmountBorrowed	Term
BidCount	0.00000008	-9.046E-05	3.4E-07	1.8E-07	0.0001282	1.9E-07	0	-4.6E-07
BorrowerRate	-9.046E-05	20.88826	-0.0041464	0.007041	-21.021193	-0.0028687	0.0000025	0.0022833
DebtToIncomeRatio	0.00000034	-0.0041464	0.0005721	-8.63E-06	0.0040276	-9.7E-07	-2E-08	-1.99E-06
IsBorrowerHomeowner	0.00000018	0.007041	-8.63E-06	0.0025895	-0.0042137	0.0000155	-5E-08	-0.0000554
LenderRate	0.00012822	-21.021193	0.0040276	-0.0042137	21.289055	0.0031878	-2.89E-06	-0.0034638
AgeInMonths	0.00000019	-0.0028687	-9.7E-07	0.0000155	0.0031878	4.87E-06	0	-6.91E-06
AmountBorrowed	0	0.0000025	-2E-08	-5E-08	-2.89E-06	0	0	0
Term	-4.6E-07	0.0022833	-1.99E-06	-0.0000554	-0.0034638	-6.91E-06	0	1.572E-05

Goodness of fit and error:

Training data			
Class	**# Cases**	**# Errors**	**% Error**
0	2986	1951	65.3
1	7013	688	9.8
Overall	9999	2639	26.4

Test data			
Class	**# Cases**	**# Errors**	**% Error**
0	3118	2031	65.1
1	7378	739	10.0
Overall	10496	2770	26.39

Exhibit E: Output from Classification Tree

Training Data scoring - Summary Report (Using Full Tree)

Classification Confusion Matrix			
	Predicted Class		
Actual Class	0	1	2
0	1242	0	1591
1	14	0	139
2	775	0	6238

Error Report			
Class	# Cases	# Errors	% Error
0	2833	1591	56.16
1	153	153	100.00
2	7013	775	11.05
Overall	9999	2519	25.19

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

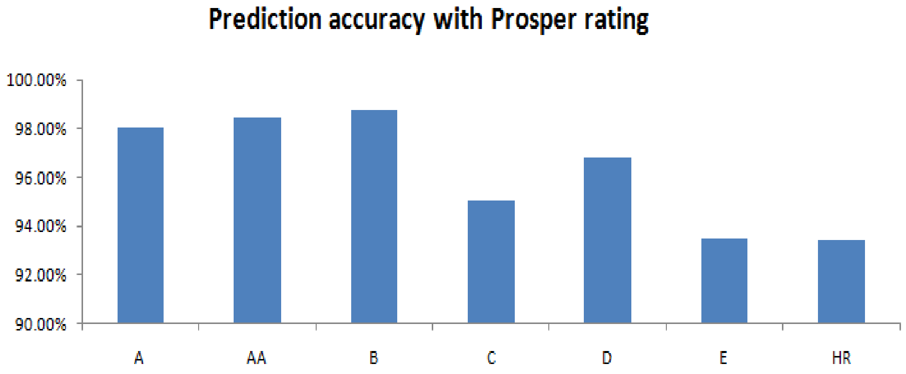
Classification Confusion Matrix			
	Predicted Class		
Actual Class	0	1	2
0	1119	0	1844
1	12	0	143
2	698	0	6680

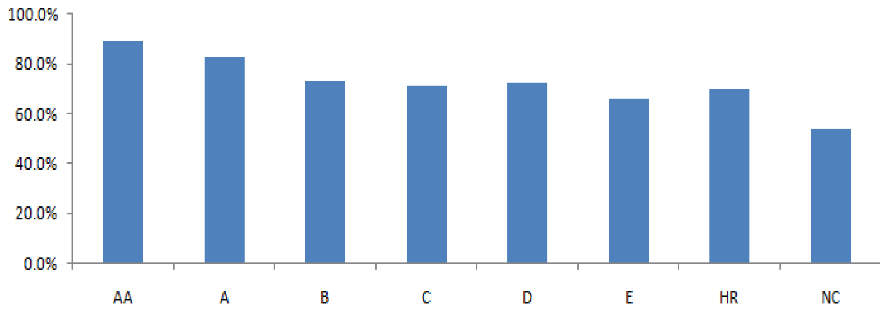
Error Report			
Class	# Cases	# Errors	% Error
0	2963	1844	62.23
1	155	155	100.00
2	7378	698	9.46
Overall	10496	2697	25.70

Errors broken down by Prosper Rating

Count of new	Error			%
New Rating	0	1	Grand Total	Accuracy
AA	138	1133	1271	89.1%
A	241	1136	1377	82.5%
B	393	1047	1440	72.7%
C	554	1349	1903	70.9%
D	531	1379	1910	72.2%
E	407	788	1195	65.9%
HR	409	939	1348	69.7%
NC	24	28	52	53.8%
Grand Total	2697	7799	10496	

Exhibit F: Application of model





Prediction accuracy with Prosper & Credit rating