WEB APPENDIX

When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications

Table of Contents

1. Additional Information about our Analyses 2

A. Procedure for coding the profile pictures 2

B. Random Forest and Extra Trees 2

C. L1 regularization regression - predictive results 3

D. Latent Dirichlet allocation (LDA) - predictive results 4

E. Linguistic Inquiry and Word Count (LIWC) - predictive results 4

2. Additional Tables and Figures 5

Table A1: Correspondence between Prosper's credit grades and FICO scores 5

Table A2: Distribution of credit grades in our sample and in the population 5

Table A3: Area under the curve (AUC) for different models and different values credit grade and word frequency 6

Table A4: Confusion matrix for loan funding versus loans recommended for funding based on our model 7

Table A5: L1 regularization binary logistic regression (1 = repayment) 8

Table A6a: Bi-grams that appeared frequently in repaid loans 16

Table A6b: Bi-grams that appeared more frequently in defaulted loans 19

Table A7: Top 120 variables with the highest importance in the Random Forest 22

Table A8: Summary statistics of dataset of all loan requests (n = 122,479) 23

Table A9a: Bi-grams that appeared frequently in funded loans 24

Table A9b: Bi-grams that appeared frequently in unfunded loans 28

Table A10: lists of words with the highest relevance measure for each LDA topic 31

Figure A1: Distribution of default likelihood for funded and rejected loans 33

Figure A2: LDA analysis – selecting the number of topics based on perplexity 33

1. Additional Information about our Analyses

A. Procedure for coding the profile pictures

About a third of the borrowers' profiles in our data (6,078 profiles) included at least one picture that is not a stock photo, however many pictures were not of the borrower, or included more than one person. To identify the borrower in the picture we manually coded the borrower's profile pictures, using the following process. If the picture included captions, we relied on it to identify the borrower (for example, "My lovely wife and I"). If the picture did not include captions and there was one adult in the picture, we assumed the adult in the picture was the borrower (following the procedure in Pope and Sydnor 2011). Once borrowers were identified, we recorded their gender (Female, Male, "Cannot Tell"), age (in three brackets: Young, Middle-aged, Old), and race (Caucasian, African American, Asian, Hispanic, or "Cannot Tell"). If the picture included more than one adult and there were no captions or if the picture did not include any adult (e.g., the picture included kids, pets, or a kitchen project) we could not identify the borrower and therefore defined the gender and race of that picture as "cannot tell". We augmented the age in unidentified pictures with the average age of the identified pictures with the three ages categories coded as 1, 2 and 3, respectively.
Each picture was evaluated by at least two different undergraduate student coders, who were unaware of the research objective. Cohen Kappas suggest fairly high levels of agreement across coders, gender = 0.89, race = 0.67, and age = 0.44.[1] Disagreements were resolved by an additional coder who served as the final judge, observing the rating of the previous coders.
We note that based on the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA) borrowers are not allowed to use race, age and gender to grant loans, however, because we have no way of ensuring that lenders indeed ignored these aspects, we include them in our model.

B. Random Forest and Extra Trees

Random Forest and Extremely Randomized Trees (Extra Trees) are ensemble of trees. The idea behind both models is to combine a large number of decision trees. In these models, trees are chosen to resolve misclassification of previously included trees. The Random Forest randomly draws with replacements subsets of the calibration data to fit each tree, and a random subset of features (variables) is used in each tree. In the Variance Selection Random Forest features are chosen based on a variance threshold determined by cross validation. The idea behind variance selection threshold is to remove features that do not meet certain threshold. By definition, features that have zero variance (same value in all samples) are removed (for further details see http://scikit-learn.org/stable/modules/feature_selection.html#variance-threshold). We tested the variance in the range of 0.001-0.0650 by increments of 0.00025. We find the variance to be in the range of 0.00175-0.00375 across folds.

In the Best Feature Selection Random Forest features are selected based on a χ^2 test. That is, we

[1] Because agreement across coders for age was lower, we also tested a model without this variable. Excluding the age variable did not qualitatively affect our results.

select the K-features with the highest χ^2 score (other approaches include F-values or mutual information criteria. For further details see http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature_selection.SelectKBest). We use cross-validation to determine the "optimal" value of K. We allowed K to vary between a minimum of 10 features and a maximum of half of the training features (over 500 features in our case), by increment of 50 features. We find the number of feature to be in the range of 60-260 across folds.

The Extra Trees is an extension of the Random Forest in which the cut-off point (the split) for each feature in the tree are also chosen at random (from a uniform distribution) and the best split among them is chosen (for further details see http://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html). We use the maximal and minimal value of each feature observed in the data to select the boundaries of the uninform distribution for each feature. See Due to the size of the feature space, we first apply a K-Best Feature Selection, as described above, to select the features to be included in the Extra Trees. We find the number of features to be in the range of 60-460 across folds.

For all tree-based methods, to limit over-fitting of the trees, we randomized the parameter optimization (Bergstra and Bengio 2012) using a 3-fold cross validation on the calibration data to determine the structure of the tree (e.g., number leaves, number of splits, depth of the tree, and criteria). We use a randomized parameter optimization rather an exhaustive search (or a grid search) due to the large number of variables in our model. The parameters are sampled from a distribution (uniform) over all possible parameter values. We set the ranges for the parameters that dictate the structure of the trees as follows:

- Number of leaves [1-11]
- Depth of the tree [3 - max number of features]
- Minimum sample split [2-11]
- Min sample leaf [1-11]
- Criteria for splits [Gini or Entropy]

Reference

Bergstra, James, and Yoshua Bengio (2012), "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research*, 13 (Feb), 281-305.

C. L1 regularization regression - predictive results

To test whether the naïve Bayes findings are sensitive to the inclusion of demographics and financial information and the interdependence among words we employ a logistic regression with an L1 penalization with same 1,052 bi-grams used in the ensemble learning and naïve Bayes analysis as well as the demographic and financial information. This analysis, while less easily interpretable than the naïve Bayes, provided very similar qualitative results (see Tables A5 and A6). The correlation between the results of the naïve Bayes and the L1 regression is 0.582 (P-value < 0.01). The L1 regression results confirm that the writing styles and intentions we identified through the naïve Bayes analysis are not merely a proxy of the demographic and financial information.

D. Latent Dirichlet allocation (LDA) - predictive results

Although the purpose of the LDA analysis was to learn about the topics discussed in loan requests rather than to predict default, we nevertheless tested the predictive ability of the uncovered topics. We find that the model that includes the LDA topics fits the data better than a model that does not include the textual information in terms of the Akaike information criterion (AIC_{LDA} = 22,242 and AIC_{notext} = 22,443). Furthermore, the likelihood ratio test significantly supports the model with the textual information relative to the model without the textual information ($LR_{DF=12}$ = 222.95, $p < 0.001$). We ran a 10-fold cross validation similar to the one conducted for the ensemble learning model. We find that the model with the LDA topics and the other textual variables (e.g., number of characters in the loan request) predicts defaults better than a baseline model that includes all the financial and demographic information but no textual information (AUC_{LDA} = 70.82% vs. AUC_{noLDA} = 70.1%). The model with the LDA variables provided higher AUC relative to the model without the textual information in all 10 folds.

E. Linguistic Inquiry and Word Count (LIWC) - predictive results

We find that the model that includes the LIWC dictionaries fits the data better than a model that does not include the textual information in terms of the Akaike information criterion (AIC_{text} = 22,250 and AIC_{notext} = 22,443), and the likelihood ratio test ($LR_{DF=69}$ = 331.54, $p < 0.001$). To test for the predictive ability of this model we ran a 10-fold cross validation similar to the one conducted for the ensemble learning model. We find that the model with LIWC predicts defaults better than a baseline model that includes all the financial and demographic information but no textual information in all 10-folds (average AUC_{LIWC} = 70.9% vs. AUC_{noLIWC} = 70.1%).

2. Additional Tables and Figures

Table A1: Correspondence between Prosper's credit grades and FICO scores

Grade	**AA**	**A**	**B**	**C**	**D**	**E**	**HR**
Score	760+	720-759	680-719	640-679	600-639	560-599	520-559

Table A2: Distribution of credit grades in our sample and in the population

FICO Score	**Prosper Credit Grade**	**Borrowers whose loans were funded**	**Distribution in the US population (Source: FICO.com)**
520-559	HR	8.1%	7.5%
560-599	E	8.4%	8.3%
600-639	D	18.2%	8.8%
640-679	C	21.5%	10.5%
680-719	B	17.4%	12.3%
720-759	A	13.4%	14.7%
760+	AA	13.1%	37.3%
Sum		100%	100%

Table A3: Area under the curve (AUC) for different models and different values credit grade and word frequency

The following is the AUC for each the five models in the ensemble with text only, financial and demographics information only, and a combination of both, for different slices of the data

	(1) Text only	(2) Financial/ demo	(3) Text & financial/demo
AUC of the underlying models of the ensemble			
Low Credit Grade (HR, E, D)			
Logistic L1	60.85%	60.90%	63.92%
Logistic L2	61.61%	58.63%	64.39%
Random Forest (Variance Selection)	59.65%	61.58%	65.21%
Random Forest (Best Features Selection)	60.42%	61.45%	63.96%
Extremely Randomized Trees (Extra Trees)	60.51%	61.87%	64.52%
Medium Credit Grade (B,C)			
Logistic L1	61.76%	65.50%	67.04%
Logistic L2	62.92%	63.65%	67.65%
Random Forest (Variance Selection)	59.81%	65.21%	65.87%
Random Forest (Best Features Selection)	60.93%	63.95%	66.46%
Extremely Randomized Trees (Extra Trees)	60.58%	64.52%	66.33%
High Credit Grade (AA, A)			
Logistic L1	71.02%	75.89%	77.90%
Logistic L2	72.04%	74.16%	77.81%
Random Forest (Variance Selection)	66.69%	77.14%	77.32%
Random Forest (Best Features Selection)	68.96%	75.65%	77.71%
Extremely Randomized Trees (Extra Trees)	69.45%	76.41%	76.86%
AUC of the underlying models of the ensemble			
Infrequent Words (Bottom 500 words)			
Logistic L1	65.39%	70.09%	70.40%
Logistic L2	66.07%	68.54%	70.79%
Random Forest (Variance Selection)	64.03%	70.22%	70.75%
Random Forest (Best Features Selection)	63.11%	69.32%	71.09%
Extremely Randomized Trees (Extra Trees)	64.91%	69.62%	70.52%
Frequent Words (Top 552 words)			
Logistic L1	67.39%	70.09%	71.66%
Logistic L2	67.47%	68.54%	71.94%
Random Forest (Variance Selection)	64.33%	70.24%	70.83%
Random Forest (Best Features Selection)	65.98%	69.30%	71.43%
Extremely Randomized Trees (Extra Trees)	66.39%	69.84%	70.95%

Notes: all AUCs are averaged across 10-folds.

Table A4: Confusion matrix for loan funding versus loans recommended for funding based on our model

In the following table we compare actual loan funding with recommended loan funding based expected profits

	Recommend based on expected profits		
Actual	Funded	Not Funded	Overall
Funded	11,795	7,651	19,446
Not Funded	21,631	81,402	103,033
Overall	33,426	89,053	122,479

Table A5: L1 regularization binary logistic regression (1 = repayment).

Results for variables with β ≠ 0

Variable	Beta	Variable	Beta	Variable	Beta
Amount Requested(x 1000)	-0.06451	year ago	1.7324	Big	1.0329
Credit Grade HR	-0.7062	health	1.7194	incom ratio	0.9883
Credit Grade E	-0.3598	side	1.7042	Purchas	0.9863
Credit Grade D	-0.2897	prosper lender	1.6078	car insur	0.9748
Credit Grade C	-0.1395	com	1.5438	Electr	0.9721
Credit Grade A	0.7631	borrow	1.5159	off the	0.9671
Credit Grade AA	0.2699	few month	1.4526	Minimum	0.9628
Group membership	-0.1045	than	1.4252	pay for	0.9581
Debt to income missing	-0.2461	and plan	1.3999	Almost	0.9552
Debt to income ratio	-0.0820	pay thi	1.3827	Active	0.9508
Images	0.0058	card debt	1.3793	that can	0.9348
Is vbrrower homeowner	-0.3090	lend	1.3540	payment and	0.9265
Lender rate	-5.4153	bonu	1.3176	the other	0.9238
New England	0.0973	dure	1.2785	Earli	0.9157
Middle East	0.2923	and our	1.2750	Larg	0.9094
Great Lakes	0.0734	unfortun	1.2689	and had	0.9055
Plains Regions	0.0640	again for	1.2634	Consult	0.8845
South West	0.0423	student loan	1.2252	Creat	0.8762
Rocky Mountain	0.2861	step	1.2135	Understand	0.8751
Far West	0.0577	reflect	1.1965	thi debt	0.8678
Military	1.3459	card with	1.1906	and current	0.8629
# number of words in description	-0.0012	goe	1.1866	Coupl	0.8593
Spelling mistakes	-0.0030	wed	1.1769	Contribut	0.8419
SMOG	0.0253	graduat	1.1619	improv credit	0.8199
% Greater than or equal to 6	-0.6424	loan payment	1.1421	the debt	0.8132
Gender male	0.0836	your	1.1342	Run	0.8053
Gender female	-0.0204	save	1.1315	they are	0.8042
Age	-0.2002	off thi	1.1286	and are	0.7835
Race white	0.1270	averag	1.1247	job with	0.7802
Race african American	-0.2140	and get	1.1161	decid	0.7780
Race asian	0.4386	fall	1.1070	colleg	0.7655

Variable	Beta	Variable	Beta	Variable	Beta
Race hispanics	0.0000	car payment	1.1004	good job	0.7631
priorListings	0.0030	grow	1.0937	along	0.7628
# of words in title	-0.0031	ani question	1.0858	cover the	0.7620
the balanc	2.0717	anoth	1.0821	past year	0.7609
august	2.0360	the cost	1.0624	owner	0.7528
invest	1.8714	the credit	1.0577	while	0.7504
reinvest	1.8078	but the	1.0485	even	0.7469
lower interest	1.7650	last year	1.0333	detail	0.7462
last	0.7447	futur	0.5782	budget	0.4155
payment for	0.7393	payment thi	0.5764	prior	0.3967
appli	0.7384	avail	0.5756	everi month	0.3964
the first	0.7382	share	0.5679	togeth	0.3922
risk	0.7353	rental	0.5674	rebuild	0.3881
ever	0.7304	have great	0.5644	there	0.3880
quickli	0.7143	return	0.5644	learn	0.3860
the payment	0.7100	have two	0.5634	through	0.3846
but have	0.7082	car loan	0.5633	max	0.3817
bank	0.7080	your consider	0.5541	experi	0.3726
student	0.7050	water	0.5479	have steadi	0.3722
over the	0.6975	stabl	0.5455	should	0.3711
although	0.6910	realiz	0.5381	for over	0.3632
you will	0.6867	least	0.5204	loan that	0.3581
low	0.6832	teach	0.5118	work the	0.3573
mistak	0.6831	sinc	0.5083	posit	0.3558
tax	0.6820	each	0.5075	reliabl	0.3506
though	0.6747	see	0.5046	process	0.3477
problem	0.6623	salari	0.4944	solid	0.3475
thank for	0.6600	never	0.4927	major	0.3456
longer	0.6597	turn	0.4885	happen	0.3337
order	0.6573	inform	0.4857	year monthli	0.3267
part	0.6523	off with	0.4800	have not	0.3217
debt free	0.6449	with prosper	0.4765	collect	0.3140
teacher	0.6409	loan from	0.4739	the bank	0.3130

Variable	Beta	Variable	Beta	Variable	Beta
the minimum	0.6408	elimin	0.4730	year now	0.3120
the high	0.6357	mean	0.4693	cover	0.3115
earn	0.6341	make payment	0.4657	expect	0.3067
use credit	0.6325	excel credit	0.4655	than the	0.3064
financ	0.6290	our credit	0.4652	life	0.3045
get out	0.6263	manag	0.4522	interest credit	0.3039
month have	0.6250	way	0.4485	close	0.3023
everi	0.6234	those	0.4478	interest rate	0.2997
could	0.6221	free	0.4401	both	0.2988
abov	0.6168	account	0.4262	june	0.2985
year have	0.6038	did not	0.4260	myself	0.2813
been pay	0.6021	time have	0.4211	the process	0.2808
rather	0.5915	file	0.4209	comput	0.2776
too	0.5823	travel	0.4205	own home	0.2762
prosper and	0.5823	credit score	0.4196	husband and	0.2739
less	0.5802	car and	0.4161	summer	0.2721
sure	0.2693	schedul	0.1596	can see	0.0362
ga	0.2688	system	0.1546	provid	0.0356
point	0.2683	payoff	0.1532	annual	0.0343
into	0.2680	entir	0.1530	cost	0.0336
clear	0.2652	made	0.1482	remain	0.0314
under	0.2637	into one	0.1472	have veri	0.0296
singl	0.2596	bankruptci	0.1458	the busi	0.0282
toward	0.2536	instead	0.1414	howev	0.0266
final	0.2525	marri	0.1414	until	0.0255
misc	0.2522	help with	0.1410	set	0.0231
cours	0.2495	littl	0.1355	build	0.0224
recent	0.2477	and not	0.1273	career	0.0214
engin	0.2470	success	0.1230	off credit	0.0205
paid for	0.2396	it	0.1229	current employ	0.0167
addit	0.2338	given	0.1190	profil	0.0138
fee	0.2335	half	0.1175	compani and	0.0096
been with	0.2305	note	0.1092	firm	0.0090

Variable	Beta	Variable	Beta	Variable	Beta
and they	0.2291	fund	0.1087	replac	0.0067
next year	0.2275	consid	0.1052	the past	0.0029
default	0.2261	profession	0.1012	been employ	0.0017
continu	0.2191	promot	0.1003	real	-0.0011
balanc	0.2165	small	0.0931	total	-0.0043
five	0.2140	time for	0.0923	gross	-0.0128
improv	0.2062	appreci	0.0900	school and	-0.0133
delinqu	0.2048	cash flow	0.0888	cell phone	-0.0141
still	0.2009	miss payment	0.0885	state	-0.0150
well	0.1978	would like	0.0874	wait	-0.0175
except	0.1972	chang	0.0801	leas	-0.0179
truck	0.1942	live	0.0790	work and	-0.0202
debt that	0.1899	look	0.0779	the purpos	-0.0217
paid off	0.1899	establish	0.0774	school	-0.0226
guarante	0.1880	degre	0.0759	and for	-0.0239
loan thank	0.1836	becaus have	0.0742	wife and	-0.0241
anyth	0.1833	extra	0.0692	came	-0.0296
thi will	0.1792	fix	0.0619	get the	-0.0301
incur	0.1673	off and	0.0594	leav	-0.0357
extrem	0.1668	self	0.0569	wa not	-0.0375
retir	0.1656	after	0.0557	record	-0.0378
offer	0.1609	ad	0.0530	most	-0.0387
payment the	0.1597	help get	0.0418	that ha	-0.0397
for our	-0.0405	plu	-0.1432	quit	-0.2260
and credit	-0.0411	rent	-0.1436	loan which	-0.2304
univers	-0.0432	thing	-0.1446	father	-0.2332
open	-0.0445	increas	-0.1450	higher	-0.2354
that wa	-0.0461	will also	-0.1484	level	-0.2397
found	-0.0484	and hope	-0.1491	higher interest	-0.2468
high interest	-0.0486	also have	-0.1508	perfect	-0.2470
ask for	-0.0492	book	-0.1554	histori	-0.2522
normal	-0.0546	equip	-0.1570	own	-0.2539
bought	-0.0555	loan becausei	-0.1583	finish	-0.2613

Variable	Beta	Variable	Beta	Variable	Beta
seek	-0.0562	year old	-0.1683	area	-0.2626
past	-0.0574	new	-0.1756	the end	-0.2644
credit rate	-0.0647	wife	-0.1771	relist	-0.2657
credit report	-0.0667	work with	-0.1782	and the	-0.2664
and can	-0.0681	pleas	-0.1806	veri respons	-0.2688
deposit	-0.0709	may	-0.1823	the time	-0.2724
work for	-0.0720	use the	-0.1828	day	-0.2770
respons	-0.0741	not have	-0.1832	will make	-0.2801
and that	-0.0792	clean	-0.1835	alway	-0.2850
score and	-0.0848	age	-0.1843	save and	-0.2853
plan	-0.0874	stand	-0.1889	credit and	-0.2865
for take	-0.0874	two year	-0.1889	love	-0.2876
individu	-0.0881	home and	-0.1905	mine	-0.2883
licens	-0.0934	the interest	-0.1931	repair	-0.2894
due	-0.0978	find	-0.1963	loan have	-0.2908
that have	-0.1032	yr	-0.1983	year with	-0.2958
live with	-0.1037	budget mortgag	-0.1984	the loan	-0.2981
onc	-0.1058	compani	-0.2020	know	-0.2982
grade	-0.1098	mother	-0.2027	obtain	-0.2989
profit	-0.1142	commun	-0.2029	dont	-0.3097
commit	-0.1150	etc	-0.2040	develop	-0.3111
you for	-0.1168	use for	-0.2061	and just	-0.3120
done	-0.1180	can pay	-0.2072	veri hard	-0.3152
need thi	-0.1235	much	-0.2144	and will	-0.3174
keep	-0.1289	off all	-0.2148	the same	-0.3217
oblig	-0.1323	do	-0.2161	attend	-0.3222
loan would	-0.1326	the money	-0.2185	room	-0.3255
sold	-0.1353	per month	-0.2204	abl	-0.3309
use thi	-0.1377	product	-0.2236	you are	-0.3351
month monthli	-0.1427	have good	-0.2252	month for	-0.3354
pay back	-0.3402	rate and	-0.4552	citi	-0.6057
pleas help	-0.3430	save for	-0.4561	assist	-0.6076

Variable	Beta	Variable	Beta	Variable	Beta
becaus the	-0.3509	receiv	-0.4622	thi prosper	-0.6096
for loan	-0.3519	request	-0.4636	call	-0.6196
dream	-0.3540	be	-0.4642	and help	-0.6275
loan pay	-0.3568	prioriti	-0.4695	between	-0.6293
and need	-0.3604	expand	-0.4734	dollar	-0.6305
deduct	-0.3608	prosper loan	-0.4745	mani	-0.6305
consolid	-0.3746	will pay	-0.4755	their	-0.6320
support	-0.3751	per	-0.4758	item	-0.6447
list and	-0.3854	these	-0.4785	then	-0.6488
surgeri	-0.3872	line	-0.4826	around	-0.6524
time job	-0.3918	charg	-0.4863	medic	-0.6612
debt and	-0.3936	are not	-0.4870	have alway	-0.6634
husband	-0.4036	properti	-0.4951	come	-0.6700
what you	-0.4074	mortgag	-0.5004	explain	-0.6767
thi time	-0.4078	within the	-0.5077	capit	-0.6855
expens car	-0.4121	stress	-0.5216	everyth	-0.6911
give	-0.4123	stabl job	-0.5249	drive	-0.6911
the next	-0.4171	report	-0.5270	have work	-0.6929
care	-0.4181	incom and	-0.5276	interest loan	-0.6980
repay thi	-0.4185	loan with	-0.5303	sever	-0.7000
from the	-0.4195	field	-0.5341	left over	-0.7007
look for	-0.4228	score	-0.5367	locat	-0.7025
afford	-0.4229	for and	-0.5378	she	-0.7078
ago and	-0.4243	which will	-0.5393	ani	-0.7148
time monthli	-0.4287	the compani	-0.5405	where	-0.7192
juli	-0.4345	worker	-0.5456	famili	-0.7302
child	-0.4383	answer	-0.5473	that need	-0.7326
and would	-0.4389	abil	-0.5525	advanc	-0.7347
help pay	-0.4391	educ	-0.5628	doe	-0.7356
date	-0.4402	tri	-0.5701	off high	-0.7390
card that	-0.4445	seem	-0.5726	the fund	-0.7458
for almost	-0.4450	thi year	-0.5766	late	-0.7624
dti	-0.4452	make the	-0.5767	behind	-0.7648

Variable	Beta	Variable	Beta	Variable	Beta
www	-0.4460	were	-0.5780	taken	-0.7650
store	-0.4469	equiti	-0.5821	difficult	-0.7653
three	-0.4543	monthli payment	-0.5947	forward	-0.7749
payment other	-0.4550	kid	-0.6002	valu	-0.7763
just need	-0.7777	long	-0.8919	websit	-1.1693
off some	-0.7845	need the	-0.9440	promis	-1.1840
sourc	-0.7865	busi	-0.9504	took	-1.1876
loan and	-0.8118	them	-1.0029	and veri	-1.1896
someon	-0.8248	refin	-1.0072	industri	-1.2068
becausei	-0.8277	bit	-1.0113	maintain	-1.2239
back thi	-0.8384	monthli incom	-1.0238	person	-1.2278
loan off	-0.8463	sale	-1.0447	daughter	-1.2799
again	-0.8499	bill and	-1.0486	fact	-1.3150
they	-0.8550	bid	-1.0486	get back	-1.3578
total monthli	-0.8579	project	-1.0511	gener	-1.3681
price	-0.8587	been the	-1.0655	follow	-1.3771
divorc	-0.8600	payday loan	-1.0711	son	-1.4448
verifi	-0.8632	hard	-1.0984	god	-1.7250
with the	-0.8642	will have	-1.1039	estat	-2.0248
the year	-0.8777	the opportun	-1.1084	lost	-2.1825
need help	-0.8899	local	-1.1657	thank you	-2.3893
				Intercept	1.9471

The table above reports the variables in the L1 regularized regression that were not set to zero. Below we list the variables that were set to zero.
Note, that while one can use bootstrap approach to obtain standard errors for the L1 regularization binary logistic regression parameter estimates, because the parameters of the L1 regularization model are biased, standard errors in a regularized regression are not meaningful (Park and Casella 2008). Accordingly, we do not report standard errors in the Table A5.

Variables with β = 0:

Bi-grams (listed here alphabetically):
abl pay, about month, about year, account and, actual, add, after tax, ago, ahead, all bill, all credit, all debt, all our, all the, allow, almost year, alreadi, alway paid, alway pay, america, amount, and also, and ha, and i'm, and make, and now, and pay, and start, and take, and thank, and then, and thi, and wa, and want, and work, apart, approx, approxim, are good, are paid, ask,

auto, automat, away, back the, back track, bad, base, becom, been late, been work, befor, begin, believ, below, benefit, best, better, bill time, bless, bring, busi and, buy, can get, can't, card balanc, card financi, card have, case, cash, catch, caus, cell, chanc, check, child support, children, class, client, combin, compani for, complet, consider, consolid credit, contact, contract, credit histori, current have, current work, custom, cut, deal, debt financi, debt have, debt incom, decis, depend, did, didn't, differ, direct, doe not, don't, don't have, down, down the, due the, dure the, each month, easili, emerg, employ, employ for, employe, end, enjoy, enough, everyon, excel, exist, expens and, expens are, expens for, expens ga, expens total, explain what, explain whi, far, feel, feel free, few, few year, figur, first, flow, for consid, for financi, for month, for pay, for prosper, for view, for year, for your, four, friend, from prosper, full, full time, fulli, further, ga util, get rid, get thi, go, goal, god bless, gone, good credit, got, great, greatli, groceri, group, ha been, happi, hard work, have alreadi, have ani, have credit, have excel, have had, have learn, have made, have never, have one, have over, have paid, have problem, have some, have stabl, have the, hello, help out, her, here, hi, him, hold, honest, hope, hospit, hour, hous and, how, i'd, i'll, i'm, i'm not, i'v, i'v been, immedi, import, includ, incom after, incom from, intend, into the, issu, it', job and, job for, know that, late payment, law, left, lender, less than, lesson, let, like pay, limit, list, loan back, loan consolid, loan credit, loan explain, loan financi, loan for, loan help, loan monthli, loan need, loan request, loan the, lot, lower, market, medic bill, meet, member, minimum payment, miss, mom, money and, money for, money pay, month ago, month and, month that, monthli budget, more than, mortgag rent, move, name, need pay, never been, never miss, next, not includ, not onli, now and, now have, number, off debt, offic, old, one payment, one the, onli, oper, opportun, origin, our, our home, out the, outstand, over year, overtim, owe, paid full, parent, part time, pass, pay all, pay bill, pay down, pay the, pay them, paycheck, payday, payment have, payment prosper, payment time, payment will, peopl, period, person loan, pictur, place, plan pay, possibl, post, present, pretti, previou, program, prosper payment, prosper will, prove, public, purpos thi, put, question, rais, ratio, read, readi, real estat, realli, reason, rebuild credit, reduc, remov, rent insur, repay, requir, rest, result, review, revolv, rid, right, right now, same, say, second, secur, see have, sell, servic, short, show, site, situat explain, situat have, six, some credit, someth, soon, spend, start, stay, steadi, strong, such, take care, take the, term, that are, that the, that thi, that time, that will, that would, that you, the amount, the best, the bill, the follow, the futur, the hous, the last, the monthli, the mortgag, the new, the onli, the prosper, the reason, the remain, the rest, there are, thi money, think, thought, three year, time and, time everi, time the, top, top prioriti, total expens, track, tri get, tuition, two, unexpect, use consolid, use help, use pay, usual, vehicl, veri good, view, view list, want, want pay, week, went, what, when, when wa, whi, whi you, who, wife', will abl, will allow, will help, will not, will paid, with credit, with thi, within, without, won't, wonder, work full, work hard, worth, would have, would use, year and, year the, yet, you can, you have, young, your help, your time

Financial and demographic variables:
Bank draft fee Annual rate, Credit Grade B, South East, Gender Unknown, Race Unknown, Race - Hispanics

Reference

Park, Trevor and George Casella (2008), “The Bayesian Lasso,” *Journal of the American Statistical Association*, 103 (482), 681-686.

Table A6a: Bi-grams that appeared frequently in repaid loans

p(word|repaid)/p(word|defaulted) ≥ 1.1

Bi-gram (repaid)	Ratio	Bi-gram (repaid)	Ratio	Bi-gram (repaid)	Ratio
reinvest	3.92	bonu	1.43	incur	1.29
lend	2.19	low	1.41	mean	1.29
lower interest	1.99	car and	1.41	everi month	1.29
i'd	1.96	off thi	1.41	earn	1.29
side	1.94	than	1.40	pretti	1.28
excel credit	1.82	cover the	1.40	and current	1.28
borrow	1.80	earli	1.39	been pay	1.28
wed	1.80	miss payment	1.38	worth	1.28
prosper lender	1.78	job with	1.38	less than	1.28
student loan	1.78	share	1.38	debt financi	1.28
than the	1.74	activ	1.38	pay for	1.27
invest	1.71	incom ratio	1.37	car insur	1.27
graduat	1.69	august	1.37	less	1.26
rather	1.68	com	1.37	ever	1.26
student	1.67	big	1.36	apart	1.26
card with	1.66	own home	1.36	default	1.26
the minimum	1.64	interest rate	1.36	health	1.26
the balanc	1.61	don't	1.35	all bill	1.26
contribut	1.59	usual	1.34	instead	1.26
it'	1.58	travel	1.34	excel	1.26
thi debt	1.56	i'm	1.34	debt have	1.26
risk	1.54	colleg	1.33	good credit	1.26
summer	1.54	guarante	1.33	miss	1.25
i'll	1.53	like pay	1.33	payment for	1.25
engin	1.53	more than	1.33	solid	1.25
card debt	1.52	easili	1.32	retir	1.25
and i'm	1.52	spend	1.32	possibl	1.25
have excel	1.52	use credit	1.32	save for	1.25
the bank	1.52	firm	1.32	debt free	1.25
and plan	1.51	paid for	1.31	understand	1.25
thank for	1.50	figur	1.31	but the	1.25
after tax	1.50	expens are	1.31	the debt	1.25
prosper and	1.50	didn't	1.31	cover	1.25
i'v been	1.50	debt incom	1.30	debt that	1.24
goe	1.50	balanc	1.30	teach	1.24
minimum payment	1.49	quickli	1.30	off credit	1.24
never miss	1.48	return	1.30	higher	1.24
i'v	1.47	rate and	1.30	and want	1.24
minimum	1.46	the interest	1.30	have steadi	1.24
entir	1.46	save	1.30	good job	1.24
the cost	1.44	expect	1.29	have great	1.24

Bi-gram (repaid)	Ratio	Bi-gram (repaid)	Ratio	Bi-gram (repaid)	Ratio
down the	1.24	think	1.20	law	1.16
salari	1.24	the first	1.20	outstand	1.16
interest credit	1.24	book	1.20	fix	1.16
alway pay	1.24	case	1.19	under	1.16
have never	1.24	decid	1.19	ad	1.16
stabl job	1.24	with credit	1.19	cost	1.16
consult	1.23	credit histori	1.19	annual	1.16
current have	1.23	dure the	1.19	coupl	1.16
lender	1.23	cours	1.19	site	1.16
stabl	1.23	pay down	1.19	ga	1.16
step	1.23	you have	1.19	past year	1.16
card balanc	1.23	toward	1.19	payment have	1.16
while	1.23	way	1.19	appli	1.16
few month	1.23	major	1.19	purchas	1.16
becaus have	1.22	ani question	1.19	comput	1.16
the credit	1.22	next year	1.18	use pay	1.16
lower	1.22	card financi	1.18	off debt	1.15
have veri	1.22	plan	1.18	car loan	1.15
payment and	1.22	averag	1.18	month have	1.15
pay thi	1.22	payment the	1.18	question	1.15
the high	1.22	tax	1.18	use consolid	1.15
teacher	1.22	term	1.18	anyth	1.15
card have	1.22	ga util	1.18	longer	1.15
plan pay	1.22	paid full	1.18	note	1.15
situat have	1.22	year monthli	1.17	the other	1.15
time for	1.22	free	1.17	profession	1.15
too	1.22	thought	1.17	have two	1.15
incom after	1.22	alreadi	1.17	buy	1.15
promot	1.22	three year	1.17	current employ	1.15
except	1.22	replac	1.17	how	1.15
though	1.22	order	1.17	reflect	1.15
univers	1.22	half	1.17	requir	1.15
about month	1.22	dure	1.17	young	1.14
have alreadi	1.21	ratio	1.17	far	1.14
have good	1.21	revolv	1.17	through	1.14
even	1.21	credit rate	1.17	parent	1.14
late payment	1.21	degre	1.17	togeth	1.14
schedul	1.21	the remain	1.17	and are	1.14
least	1.21	don't have	1.17	extra	1.14
one the	1.21	futur	1.17	max	1.14
higher interest	1.21	fall	1.17	would like	1.14
your consider	1.21	happi	1.17	live	1.14
expens ga	1.20	have credit	1.17	thi money	1.13
below	1.20	veri good	1.17	paid off	1.13

Bi-gram (repaid)	Ratio	Bi-gram (repaid)	Ratio	Bi-gram (repaid)	Ratio
look	1.13	post	1.12	base	1.11
loan the	1.13	have ani	1.12	june	1.10
histori	1.13	are paid	1.12	paycheck	1.10
last year	1.13	someth	1.12	and pay	1.10
offer	1.13	both	1.12	yet	1.10
bit	1.13	part time	1.11	although	1.10
have stabl	1.13	class	1.11	time have	1.10
consolid credit	1.13	few year	1.11	group	1.10
fulli	1.13	loan from	1.11	make payment	1.10
strong	1.13	further	1.11	full time	1.10
into one	1.13	over the	1.11	career	1.10
addit	1.13	reason	1.11	work full	1.10
financ	1.13	that can	1.11	cash flow	1.10
off high	1.13	charg	1.11	money and	1.10
and would	1.12	larg	1.11	and hope	1.10
profil	1.12	out the	1.11	not includ	1.10
reliabl	1.12	reduc	1.11	the last	1.10
five	1.12	should	1.11	tuition	1.10
best	1.12	short	1.11	limit	1.10
year ago	1.12	intend	1.11	the monthli	1.10
experi	1.12	card that	1.11	sure	1.10

Table A6b: Bi-grams that appeared more frequently in defaulted loans

p(word|defaulted)/p(word|repaid) ≥ 1.1

Bi-gram (defaulted)	Ratio	Bi-gram (defaulted)	Ratio	Bi-gram (defaulted)	Ratio
payday loan	2.12	locat	1.47	mother	1.31
payday	2.06	real estat	1.46	thi prosper	1.30
god	2.01	see have	1.46	children	1.30
god bless	2.00	estat	1.46	sale	1.30
view list	1.99	daughter	1.45	hard	1.30
need help	1.85	are good	1.44	child	1.30
for view	1.84	time everi	1.42	father	1.30
top prioriti	1.84	caus	1.42	have over	1.30
lost	1.77	pleas help	1.41	rebuild	1.30
bless	1.73	refin	1.41	rebuild credit	1.30
the follow	1.69	project	1.40	tri get	1.29
view	1.67	follow	1.40	mother	1.31
prioriti	1.67	back thi	1.40	thi prosper	1.30
promis	1.66	medic bill	1.40	children	1.30
prosper will	1.65	divorc	1.39	sale	1.30
for prosper	1.65	you are	1.39	hard	1.30
would use	1.65	left over	1.39	child	1.30
payment prosper	1.63	just need	1.39	father	1.30
list and	1.63	and credit	1.38	have over	1.30
get back	1.63	prove	1.37	rebuild	1.30
back track	1.61	veri hard	1.37	rebuild credit	1.30
behind	1.60	for pay	1.37	tri get	1.29
yr	1.59	again	1.36	month that	1.29
stress	1.58	call	1.36	will abl	1.29
loan explain	1.56	real	1.35	know that	1.29
situat explain	1.55	not onli	1.35	she	1.29
son	1.54	been the	1.35	industri	1.28
explain what	1.53	hello	1.34	store	1.28
help get	1.52	worker	1.33	automat	1.28
prosper payment	1.51	have learn	1.33	off some	1.27
someon	1.51	total monthli	1.33	relist	1.27
what you	1.50	expand	1.33	local	1.27
again for	1.50	capit	1.32	lesson	1.27
explain whi	1.50	top	1.32	assist	1.27
catch	1.50	the opportun	1.31	surgeri	1.27
child support	1.49	hard work	1.31	normal	1.26
and thank	1.49	everyon	1.31	equip	1.26
explain	1.49	fact	1.31	busi and	1.26
chanc	1.49	thank you	1.31	oper	1.26
whi you	1.47	mother	1.31	bill and	1.26

Bi-gram (defaulted)	Ratio	Bi-gram (defaulted)	Ratio	Bi-gram (defaulted)	Ratio
medic	1.26	area	1.21	everyth	1.17
pay back	1.26	honest	1.21	need the	1.17
citi	1.26	loan pay	1.21	taken	1.17
famili	1.25	report	1.21	file	1.17
day	1.25	thi time	1.20	will have	1.17
difficult	1.25	custom	1.20	mistak	1.17
support	1.25	kid	1.20	total	1.17
bad	1.25	for financi	1.20	the bill	1.17
name	1.25	and wa	1.20	payment other	1.17
item	1.25	ha been	1.20	prosper loan	1.16
the busi	1.25	repair	1.20	person	1.16
direct	1.25	hospit	1.20	person loan	1.16
husband	1.25	let	1.20	and just	1.16
advanc	1.24	attend	1.20	them	1.16
work hard	1.24	pleas	1.20	clean	1.16
year old	1.24	hi	1.20	and start	1.16
work with	1.24	verifi	1.19	interest loan	1.16
busi	1.24	that you	1.19	with the	1.16
mom	1.24	the prosper	1.19	got	1.16
took	1.24	have alway	1.19	sever	1.16
sourc	1.24	and help	1.19	mortgag rent	1.15
for take	1.24	monthli budget	1.19	her	1.15
that need	1.24	and can	1.19	budget mortgag	1.15
payment will	1.23	and need	1.19	issu	1.15
product	1.23	greatli	1.19	place	1.15
becaus the	1.23	pass	1.19	whi	1.15
ask for	1.23	you will	1.19	need thi	1.15
track	1.23	opportun	1.18	profit	1.15
you for	1.23	all our	1.18	present	1.15
our home	1.23	time monthli	1.18	open	1.15
gener	1.23	care	1.18	the mortgag	1.15
left	1.23	loan request	1.18	maintain	1.14
take care	1.22	save and	1.18	the compani	1.14
were	1.22	Old	1.18	leas	1.14
develop	1.22	licens	1.18	due the	1.14
credit report	1.22	the new	1.18	year the	1.14
websit	1.22	i'm not	1.18	age	1.14
the fund	1.22	contract	1.18	request	1.14
which will	1.22	overtim	1.18	loan for	1.14
came	1.22	Him	1.18	their	1.14

Bi-gram (defaulted)	Ratio	Bi-gram (defaulted)	Ratio	Bi-gram (defaulted)	Ratio
loan monthli	1.14	properti	1.12	help pay	1.11
List	1.14	who	1.12	Check	1.11
can get	1.14	the time	1.12	and get	1.11
Mani	1.14	properti	1.12	Perfect	1.10
These	1.14	Who	1.12	have work	1.10
Know	1.14	Where	1.12	room	1.10
Review	1.13	Market	1.12	client	1.10
off all	1.13	Gone	1.12	for year	1.10
that ha	1.13	month for	1.12	can see	1.10
Turn	1.13	payment time	1.12	give	1.10
for month	1.13	our credit	1.12	love	1.10
total expens	1.13	each month	1.11	they are	1.10
What	1.13	Juli	1.11	mortgag	1.10
Went	1.13	Obtain	1.11	doe	1.10
america	1.13	expens total	1.11	onc	1.10
the reason	1.13	Readi	1.11	self	1.10
loan and	1.13	Answer	1.11	date	1.10
deduct	1.13	Owner	1.11	deal	1.10
improv credit	1.13	back the	1.11	provid	1.10
wonder	1.13	from the	1.11	leav	1.10
begin	1.13	that wa	1.11	then	1.10
Due	1.13	and that	1.11	becausei	1.10
Our	1.13	deposit	1.11	loan becausei	1.10
been with	1.13	Remov	1.11	abil	1.10
can't	1.12	Found	1.11	all debt	1.10
and ha	1.12	that are	1.11	oblig	1.10
loan which	1.12	Afford	1.11	tri	1.10
They	1.12	and now	1.11	servic	1.10
need pay	1.12	Owe	1.11	come	1.10
the time	1.12	Mine	1.11		

Table A7: Top 120 variables with the highest importance in the Random Forest

Variables	Importance	Variables	Importance	Variables	Importance	Variables	Importance
Lender rate	0.056502	SMOG	0.003126	whi you	0.002529	look	0.002311
Credit Grade A	0.041125	get back	0.003051	abl	0.002518	Gender - female	0.002301
Credit Grade HR	0.027996	daughter	0.003031	graduat	0.002515	total	0.002292
Amount requested	0.018964	our	0.003023	never	0.002509	low	0.002280
Credit Grade E	0.012061	student loan	0.003009	famili	0.002508	bill and	0.002247
Credit Grade AA	0.010833	explain what	0.002992	will abl	0.002507	prosper loan	0.002226
Borrower homeownership	0.009590	start	0.002965	colleg	0.002505	you for	0.002223
Credit Grade D	0.007955	behind	0.002953	for year	0.002503	son	0.002221
Prior listings	0.007312	due	0.002948	gas	0.002469	pay thi	0.002218
payday loan	0.005989	interest rate	0.002919	last	0.002446	report	0.002217
Credit Grade C	0.005112	view list	0.002909	medic	0.002440	paid off	0.002210
Far West	0.005093	lend	0.002902	fund	0.002436	old	0.002208
busi	0.005073	Spell checker	0.002831	what	0.002431	list	0.002200
Middle East	0.005069	card debt	0.002821	tri	0.002412	Rocky Mountain	0.002183
borrow	0.004620	again	0.002792	loan and	0.002404	even	0.002171
invest	0.004574	Age	0.002774	live	0.002397	use thi	0.002169
than	0.004494	whi	0.002758	they	0.002394	god	0.002164
Debt to income	0.004327	them	0.002754	mortgag	0.002374	compani	0.002163
# of words in title	0.004297	balanc	0.002734	pay for	0.002370	ha been	0.002161
% words with 6 or more letters	0.004277	with the	0.002706	reinvest	0.002370	have never	0.002150
hard	0.004270	estat	0.002705	who	0.002368	the balanc	0.002148
Race - white	0.004196	what you	0.002687	and will	0.002366	these	0.002148
thank you	0.004131	you are	0.002673	real estat	0.002349	see	0.002116
person	0.004131	pay back	0.002654	into	0.002346	the time	0.002115
payday	0.004053	pleas	0.002644	more than	0.002341	know	0.002113
# of words in description	0.003848	back thi	0.002618	just need	0.002339	rather	0.002109
explain	0.003794	Race -Afr. American	0.002569	purchas	0.002335	promis	0.002103
Gender - male	0.003756	score	0.002556	total monthli	0.002334	hi	0.002093
save	0.003476	Plains Regions	0.002548	plan	0.002328	support	0.002090
student	0.003193	and the	0.002537	husband	0.002315	give	0.002085

Table A8: Summary statistics of dataset of all loan requests (n = 122,479)

Variables		Min	Max	Mean	SD	Freq.
Amount requested		1,000	25,000	7,411.1	6,189.4	
Debt-to-income ratio		0	10.01	.54	1.33	
Lender interest rate		0	.350	.196	.092	
Number of words in description		1	782	171.6	122.96	
# Prior Listings		0	67	0.90	2.06	
Credit grade:	AA					0.026
	A					0.034
	B					0.055
	C					0.105
	D					0.160
	E					0.181
	HR					0.436
Loan status (1 = Funded, 0 = Expired)						0.159
Loan image dummy						0.498
Home owner dummy						0.357

Table A9a: Bi-grams that appeared frequently in funded loans

Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio
reinvest	4.70	com	1.73	averag	1.53
relist	3.36	below	1.71	off with	1.53
prosper lender	3.36	for prosper	1.71	add	1.53
excel credit	2.54	lender	1.71	review	1.53
prosper and	2.51	consult	1.70	remain	1.53
total expens	2.47	post	1.70	properti	1.53
bid	2.46	loan thank	1.69	list and	1.52
thi prosper	2.36	loan request	1.69	entir	1.52
group	2.27	miss payment	1.68	expect	1.52
feel free	2.24	fund	1.67	cover	1.52
invest	2.22	fulli	1.67	than the	1.51
card balanc	2.21	earli	1.66	i'll	1.51
revolv	2.19	previou	1.66	solid	1.51
question	2.17	www	1.65	addit	1.50
dti	2.12	have over	1.64	prosper will	1.50
have excel	2.11	public	1.64	spend	1.50
after tax	2.10	grade	1.64	share	1.49
verifi	2.09	intend	1.64	the minimum	1.49
ani question	2.08	and plan	1.64	annual	1.49
lend	2.07	can see	1.63	base	1.48
america	2.04	prosper payment	1.62	your consider	1.48
from prosper	2.03	have never	1.62	the cost	1.48
for consid	2.03	develop	1.62	never been	1.48
cash flow	2.02	late payment	1.62	tax	1.48
prosper loan	2.01	line	1.62	price	1.48
i'd	1.96	excel	1.61	comput	1.47
flow	1.95	the balanc	1.61	request	1.47
rental	1.94	plan pay	1.61	capit	1.47
equiti	1.92	monthli incom	1.61	paid full	1.47
cover the	1.89	expens are	1.60	site	1.47
bonu	1.89	you have	1.60	you can	1.47
origin	1.87	see have	1.60	avail	1.47
list	1.87	not includ	1.59	summer	1.46
incom after	1.86	lower interest	1.58	wed	1.46
misc	1.82	delinqu	1.58	system	1.46
answer	1.81	the remain	1.58	experi	1.46
record	1.81	firm	1.58	travel	1.45
contribut	1.81	easili	1.58	and thank	1.45
thank for	1.80	figur	1.57	cost	1.45
balanc	1.80	approxim	1.57	payment thi	1.45
borrow	1.79	side	1.56	plan	1.45
the prosper	1.79	usual	1.56	for your	1.45
card with	1.77	cash	1.56	worth	1.45
project	1.77	commun	1.56	minimum payment	1.44
with prosper	1.76	profession	1.56	schedul	1.44
engin	1.75	left over	1.55	earn	1.44
default	1.75	univers	1.55	for view	1.44
gross	1.74	case	1.54	student loan	1.44
never miss	1.74	minimum	1.54	cell	1.43
rather	1.74	valu	1.53	term	1.43

Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio
again for	1.43	follow	1.34	the amount	1.28
replac	1.43	account and	1.34	incom from	1.28
ratio	1.42	salari	1.34	toward	1.28
activ	1.42	decid	1.34	sale	1.27
profil	1.41	consider	1.34	abov	1.27
save	1.41	down the	1.34	paid off	1.27
requir	1.41	detail	1.34	payment the	1.27
loan becausei	1.41	amount	1.33	thank you	1.27
loan from	1.41	june	1.33	such	1.27
gener	1.41	the bank	1.33	deduct	1.27
enjoy	1.41	graduat	1.33	client	1.27
hello	1.40	for take	1.33	from the	1.27
rate and	1.40	low	1.33	free	1.27
the first	1.40	websit	1.33	between	1.27
lower	1.40	loan credit	1.33	take the	1.27
view list	1.40	save and	1.33	the next	1.26
number	1.40	alreadi	1.32	higher interest	1.26
debt incom	1.40	loan with	1.32	ga	1.26
have ani	1.40	top prioriti	1.32	each	1.26
been late	1.39	off thi	1.32	teacher	1.26
guarante	1.39	etc	1.32	short	1.26
first	1.39	quickli	1.32	less	1.26
see	1.39	your	1.31	higher	1.26
view	1.39	than	1.31	member	1.26
you for	1.39	paid for	1.31	have veri	1.26
immedi	1.38	each month	1.31	more than	1.26
bank	1.38	book	1.31	extrem	1.25
profit	1.38	combin	1.31	account	1.25
credit histori	1.38	includ	1.31	use credit	1.25
student	1.38	juli	1.31	limit	1.25
will paid	1.38	total	1.30	histori	1.25
have credit	1.37	contact	1.30	emerg	1.25
inform	1.37	interest rate	1.30	offic	1.25
groceri	1.37	never	1.30	level	1.25
car insur	1.37	custom	1.29	last	1.25
have alreadi	1.37	over year	1.29	ani	1.25
year the	1.36	goe	1.29	thought	1.25
one the	1.36	pretti	1.29	card have	1.25
wife'	1.36	the other	1.29	expand	1.25
payment prosper	1.36	electr	1.29	strong	1.24
less than	1.36	three year	1.29	the end	1.24
here	1.35	larg	1.29	charg	1.24
alway pay	1.35	the interest	1.29	i'm not	1.24
citi	1.35	automat	1.29	teach	1.24
incom ratio	1.35	should	1.29	becausei	1.24
pay down	1.35	servic	1.29	real estat	1.24
market	1.35	pictur	1.28	the follow	1.24
note	1.35	miss	1.28	complet	1.24
cell phone	1.34	increas	1.28	oper	1.24
bought	1.34	risk	1.28	payment will	1.24
reduc	1.34	sourc	1.28	late	1.24
current have	1.34	the busi	1.28	month that	1.24

Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio
dure the	1.24	top	1.18	four	1.14
close	1.24	contract	1.18	reflect	1.14
with thi	1.24	around	1.18	fact	1.14
locat	1.24	everyon	1.18	long	1.14
estat	1.24	area	1.18	rest	1.14
local	1.23	manag	1.17	remov	1.14
offer	1.23	cours	1.17	while	1.14
save for	1.23	far	1.17	wife and	1.14
room	1.23	over the	1.17	the rest	1.14
actual	1.23	the fund	1.17	the credit	1.14
sold	1.23	industri	1.17	process	1.14
bit	1.22	promot	1.17	may	1.14
half	1.22	field	1.16	class	1.14
success	1.22	deal	1.16	month ago	1.14
auto	1.22	after	1.16	well	1.13
consid	1.22	purchas	1.16	oblig	1.13
return	1.22	month have	1.16	show	1.13
use for	1.22	those	1.16	loan which	1.13
about month	1.22	almost year	1.16	sinc	1.13
payment have	1.22	the compani	1.16	doe not	1.13
within	1.21	use the	1.16	approx	1.13
next	1.21	the year	1.16	apart	1.13
i'v	1.21	current employ	1.15	that the	1.13
i'v been	1.21	your time	1.15	month for	1.13
the monthli	1.21	year now	1.15	most	1.12
mean	1.21	quit	1.15	read	1.12
prioriti	1.21	tuition	1.15	stabl job	1.12
friend	1.21	wonder	1.15	lesson	1.12
card debt	1.21	period	1.15	same	1.12
next year	1.21	continu	1.15	appli	1.12
time everi	1.21	licens	1.15	time the	1.12
owner	1.21	sell	1.15	year have	1.12
are paid	1.21	compani for	1.15	further	1.12
about year	1.20	except	1.15	debt free	1.12
the new	1.20	compani	1.15	the last	1.12
employe	1.20	drive	1.15	the time	1.12
extra	1.20	major	1.15	expens and	1.12
leas	1.20	benefit	1.15	into the	1.12
per month	1.20	been employ	1.14	interest credit	1.12
year with	1.20	reason	1.14	and i'm	1.12
per	1.19	ad	1.14	normal	1.12
happi	1.19	build	1.14	finish	1.12
for over	1.19	card that	1.14	three	1.12
thi year	1.19	car payment	1.14	that thi	1.11
water	1.19	big	1.14	seek	1.11
real	1.19	think	1.14	dure	1.11
work the	1.19	perfect	1.14	posit	1.11
commit	1.19	thi debt	1.14	retir	1.11
veri respons	1.19	both	1.14	young	1.11
loan off	1.19	off the	1.14	everi month	1.11
grow	1.19	two year	1.14	august	1.11
refin	1.18	payment for	1.14	time have	1.11

Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio	Bi-gram (funded)	Ratio
the high	1.11	doe	1.10	their	1.10
own	1.11	elimin	1.10	store	1.10
unexpect	1.11	bill time	1.10	two	1.10
few month	1.11	product	1.10	for our	1.10
feel	1.11	good credit	1.10	prior	1.10
down	1.11	incom and	1.10	mortgag	1.10
within the	1.11	last year	1.10	coupl	1.10
result	1.11	almost	1.10	attend	1.10
the same	1.11	own home	1.10	the mortgag	1.10
possibl	1.11				

Table A9b: Bi-grams that appeared frequently in unfunded loans

Bi-gram (unfunded)	Ratio	Bi-gram (unfunded)	Ratio	Bi-gram (unfunded)	Ratio
situat explain	4.18	debt financi	1.70	and help	1.41
loan explain	4.12	budget mortgag	1.68	god	1.41
explain what	3.90	off all	1.68	stress	1.40
explain whi	3.88	help get	1.67	like pay	1.40
whi you	3.84	pleas help	1.67	that can	1.40
what you	3.74	payment other	1.63	get thi	1.40
for financi	3.69	veri hard	1.63	son	1.38
for pay	3.55	take care	1.62	into one	1.38
are good	3.50	off debt	1.62	loan pay	1.38
explain	3.49	hard	1.62	afford	1.38
you are	3.26	divorc	1.62	back the	1.38
back thi	3.09	mother	1.60	money pay	1.38
catch	3.06	need thi	1.60	get out	1.38
you will	2.58	medic bill	1.60	off credit	1.37
get back	2.54	honest	1.60	medic	1.37
back track	2.53	can't	1.59	payday	1.37
pay back	2.44	tri	1.58	job and	1.37
loan monthli	2.35	good job	1.57	loan would	1.36
someon	2.29	and just	1.56	the bill	1.36
loan for	2.28	rent insur	1.55	loan consolid	1.36
use thi	2.27	need pay	1.51	expens total	1.36
behind	2.26	and need	1.51	child support	1.36
chanc	2.24	clean	1.5	help pay	1.35
just need	2.21	can get	1.5	got	1.35
whi	2.21	ahead	1.49	have had	1.35
off some	2.01	children	1.48	famili	1.35
tri get	1.97	thing	1.48	better	1.34
worker	1.97	prove	1.48	have learn	1.34
one payment	1.97	surgeri	1.48	rebuild credit	1.34
loan need	1.90	everyth	1.47	mistak	1.34
track	1.89	mom	1.47	given	1.33
bill and	1.86	kid	1.46	singl	1.33
need help	1.84	mortgag rent	1.45	our credit	1.33
what	1.79	daughter	1.45	and want	1.32
lost	1.78	and had	1.44	debt that	1.32
bad	1.77	rebuild	1.44	clear	1.32
dont	1.73	want pay	1.43	went	1.32
and start	1.73	work hard	1.43	child	1.32
payday loan	1.73	can pay	1.43	debt and	1.32
and get	1.71	hard work	1.41	abl pay	1.32
when wa	1.31	and wa	1.23	car and	1.17

Bi-gram (unfunded)	Ratio	Bi-gram (unfunded)	Ratio	Bi-gram (unfunded)	Ratio
life	1.30	loan back	1.23	not have	1.16
give	1.30	realli	1.23	issu	1.16
help out	1.29	ask for	1.23	them	1.16
will abl	1.29	work and	1.23	could	1.16
care	1.29	all credit	1.22	the past	1.16
all debt	1.29	stand	1.22	have steadi	1.16
and now	1.29	use help	1.22	will make	1.16
and can	1.28	situat have	1.22	consolid credit	1.16
difficult	1.28	and make	1.22	year	1.15
paycheck	1.28	monthli budget	1.21	off high	1.15
will help	1.28	have made	1.21	month monthli	1.15
pay them	1.28	truck	1.21	would like	1.15
consolid	1.28	year monthli	1.21	she	1.15
turn	1.27	husband and	1.20	loan financi	1.14
hospit	1.27	improv credit	1.20	the best	1.14
start	1.27	him	1.20	sever	1.14
help with	1.27	pass	1.20	assist	1.14
seem	1.27	happen	1.20	job for	1.14
past	1.27	due	1.20	file	1.14
know that	1.26	father	1.20	and would	1.14
husband	1.26	myself	1.20	will pay	1.13
outstand	1.26	time job	1.20	and credit	1.13
caus	1.26	problem	1.19	her	1.13
want	1.26	have some	1.19	ga util	1.13
dream	1.25	all the	1.19	that have	1.13
right now	1.25	now and	1.19	say	1.13
but have	1.25	the opportun	1.19	loan payment	1.13
off and	1.24	and that	1.18	school and	1.13
know	1.24	becaus the	1.18	look for	1.13
job with	1.24	and work	1.18	find	1.13
decis	1.24	now have	1.18	all our	1.13
have one	1.24	old	1.18	and also	1.13
the purpos	1.24	repair	1.17	meet	1.13
some credit	1.24	live with	1.17	someth	1.12
work full	1.24	and pay	1.17	god bless	1.12
that are	1.24	bring	1.17	pay bill	1.12
that need	1.23	need the	1.17	card financi	1.12
all bill	1.23	gone	1.17	hi	1.12
bless	1.23	greatli	1.17	payoff	1.12

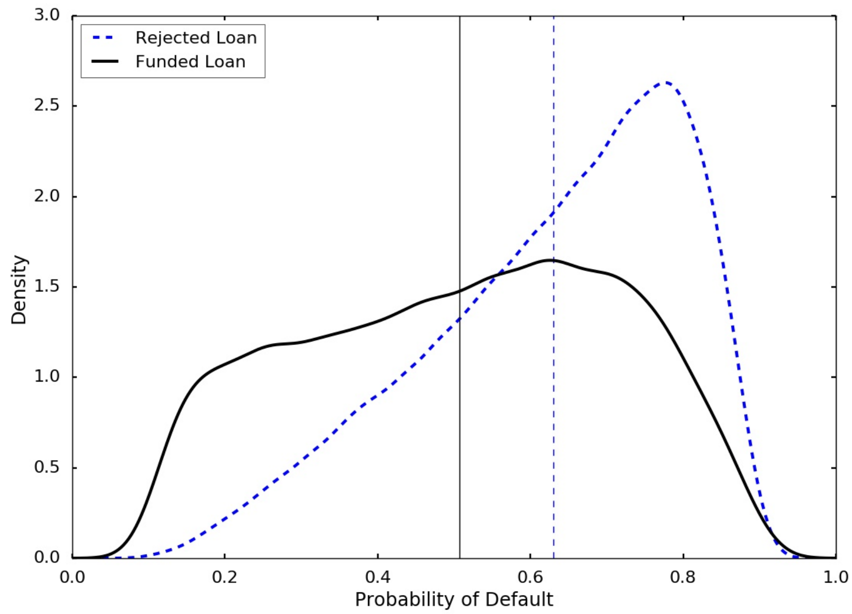
Bi-gram (unfunded)	Ratio	Bi-gram (unfunded)	Ratio	Bi-gram (unfunded)	Ratio
credit score	1.12	would have	1.11	let	1.11
order	1.12	onc	1.11	money and	1.10
work with	1.12	becom	1.11	and not	1.10
money for	1.12	obtain	1.11	loan have	1.10
direct	1.12	until	1.11	steadi	1.10
unfortun	1.12	taken	1.11	and take	1.10
pleas	1.11	and then	1.11	vehicl	1.10
that would	1.11	won't	1.11	loan the	1.10
incur	1.11	credit score	1.12	week	1.10
pay all	1.11	score	1.11		

Table A10: lists of words with the highest relevance measure for each LDA topic

Topic: Employment and School	Relevance	Topic: Interest Rate Reduction	Relevance	Topic: Expense Explanation	Relevance
work	-0.42678	debt	1.12548	expens	0.97497
job	-0.42930	interest	0.98818	explain	0.79611
full	-0.86215	rate	0.95097	cloth	0.77201
school	-0.89078	high	0.93643	entertain	0.75424
year	-1.07493	consolid	0.87551	cabl	0.75197
colleg	-1.09435	score	0.86722	whi	0.70915
incom	-1.14446	improv	0.74883	util	0.70817
employ	-1.14622	lower	0.69804	insur	0.60867
student	-1.21860	balanc	0.66615	monthli	0.39539
part	-1.30739	card	0.64182	cardsmi	0.18691
financi	-1.33920	histori	0.62272	purpos	0.17408
steadi	-1.44914	higher	0.61019	billsmi	0.15935
stabl	-1.46063	payoff	0.59608	hous	0.15739
graduat	-1.48860	reduc	0.53922	debtmi	0.10787
loan	-1.50922	elimin	0.52508	monthmonthli	0.06338
secur	-1.53972	minimum	0.51484	timemonthli	-0.00063
degre	-1.56564	outstand	0.51241	situat	-0.00922
monthli	-1.58314	low	0.50942	incomemonthli	-0.02425
educ	-1.59005	rid	0.50788	iam	-0.02954
hour	-1.60727	ratio	0.50312	loansmi	-0.04908
retir	-1.62923	goal	0.46530	card	-0.06722
finish	-1.63329	revolv	0.44719	loanmi	-0.06818
veri	-1.63949	refin	0.43334	paymentmi	-0.08156
start	-1.65941	incur	0.39694	businessmi	-0.09741
repair	-1.68777	oblig	0.37474	consolidationmi	-0.10197
thi	-1.70717	default	0.36558	loanmonthli	-0.11064
wed	-1.73917	faster	0.34038	debtsmi	-0.12716
summer	-1.76588	sooner	0.32769	incom	-0.15812
career	-1.78935	miss	0.32551	yearsmonthli	-0.16602
buy	-1.79148	apr	0.31495	cardmi	-0.16639

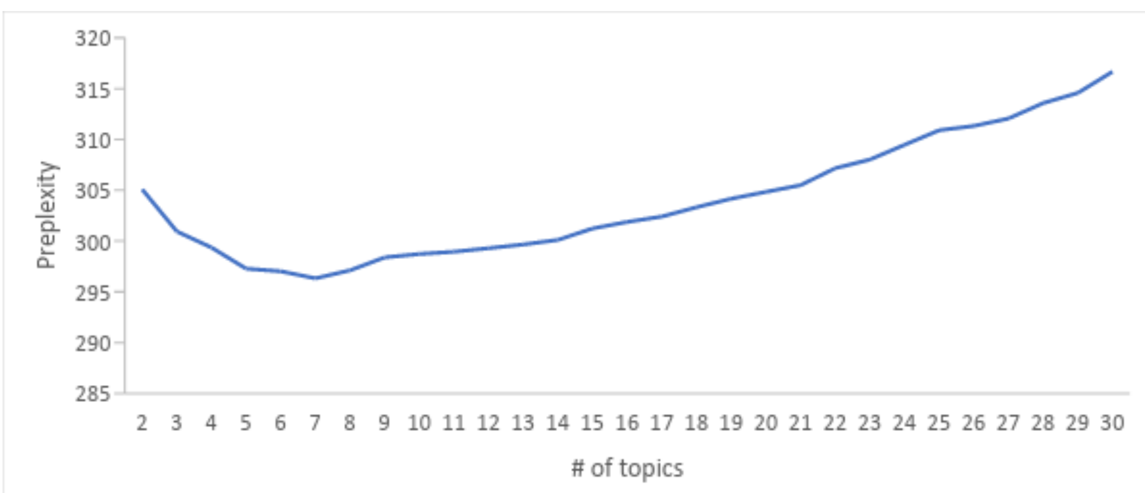
Topic: Business and Real Estate	Relevance	Topic: Family	Relevance	Topic: Loan Details and Explanations	Relevance	Topic: Monthly Payment	Relevance
busi	-0.72468	bill	-0.80362	loan	0.02223	month	-0.46588
purchas	-1.22119	tri	-0.97478	thi	-0.23886	payment	-0.62683
compani	-1.24243	famili	-1.06365	becaus	-0.39797	paid	-0.82288
invest	-1.45524	life	-1.13109	candid	-0.47236	total	-1.03183
fund	-1.49916	husband	-1.19515	situat	-0.91190	account	-1.03609
addit	-1.51557	medic	-1.21212	financi	-0.92870	rent	-1.04262
properti	-1.55064	thing	-1.29017	purpos	-0.93624	mortgag	-1.06368
market	-1.56694	littl	-1.34994	hous	-0.98446	save	-1.14949
build	-1.57312	realli	-1.35012	expens	-1.11843	list	-1.23411
cost	-1.61591	care	-1.37252	monthli	-1.12610	everi	-1.27680
sell	-1.62173	give	-1.38193	incom	-1.55843	payday	-1.29964
servic	-1.62783	children	-1.41594	entertain	-1.84658	budget	-1.32913
sale	-1.64196	hard	-1.43166	cloth	-1.90259	report	-1.33987
real	-1.71844	daughter	-1.44836	cabl	-1.92371	bank	-1.36469
alreadi	-1.74378	chanc	-1.45170	util	-1.95598	tax	-1.39207
success	-1.77930	son	-1.46681	insur	-2.02035	includ	-1.42280
open	-1.77941	money	-1.50048	bill	-2.17176	current	-1.43029
provid	-1.78311	divorc	-1.52915	card	-2.20032	amount	-1.48126
base	-1.79599	move	-1.54901	canid	-2.73682	check	-1.51827
home	-1.82519	track	-1.55675	ontim	-3.59224	delinqu	-1.55218
estat	-1.83556	everyth	-1.55791	honest	-3.98226	amp	-1.55378
profit	-1.85000	abl	-1.57058	thanksmonthli	-4.09312	onli	-1.55612
grow	-1.86056	bad	-1.57478	alway	-4.30764	day	-1.60811
offic	-1.87716	mother	-1.58898	vacat	-4.30948	left	-1.60841
run	-1.88103	home	-1.63274	buy	-4.83280	sinc	-1.69568
product	-1.88769	child	-1.64554	payback	-4.97132	bankruptci	-1.69810
store	-1.89849	kid	-1.67359	trustworthi	-5.35212	fee	-1.72871
rental	-1.90456	put	-1.67912	fix	-5.43237	auto	-1.74107
industri	-1.91004	pleas	-1.68211	catch	-6.89838	owe	-1.75320
area	-1.92412	live	-1.68560	track	-7.09870	rebuild	-1.76605

Figure A1: Distribution of default likelihood for funded and rejected loans



Note: the vertical lines are the average default probabilities.

Figure A2: LDA analysis – selecting the number of topics based on perplexity



Note: we measure perplexity as: $perplexity = -\frac{L(w)}{count\ of\ words}$, where $L(w)$ is the log-likelihood of the test data documents. Thus, perplexity is decreasing in likelihood (lower perplexity means better fit).