Peer to Peer Loan Clustering

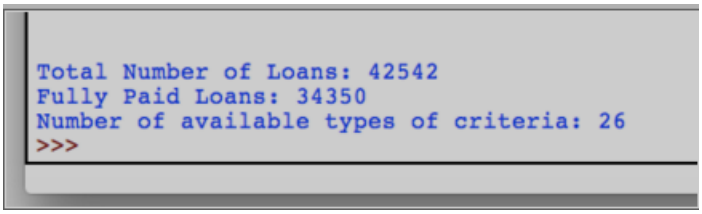
Johnathan Pasch
Joshua Kruger
Sean Ray

Project Changes: There is no major project changes at this point.

Data collection: We have collected the peer to peer lending data from LendingClub.com. While they do not offer an API, they do have logs of all pervious and ongoing loan data available for signed up "investors". Overall, each year is amounts to roughly 30MB of collected data. There are over 120 collected criteria available for us to mine. After inspecting the data, we have decided to omit data from beyond 2011 so that we don't inspect loans that have not had a chance to be repaid. Had we collected data from 2011, charge offs and defaults would have represented a higher weight in the overall collection. In the data from 2011 and before, we have collected many relevant fields, below are some of the most important which we believe will be crucial for our clustering algorithm.

Loan amount	Funded amount	Interest rate	Instalment size
Investment Grade	Employment Title	Employment length	Home owner
Annual income	Loan status	Description/Reason	Total Payment
Loan term			

We wrote a sample python program to prove that we can manipulate, and have inspected, the data. Below are key criteria from our collected data. In the future we feel it would be best to load the data into a SQL style database so that we can perform efficient queries which allow the user to cluster in real time (in the browser). On the following page there is a comprehensive list of collected criteria.



Uses of the data: The primary objective of the project is to cluster previous loans so an investor can determine where a new loan stands as far as risk, and finical return. Because of this, picking from the provided criteria will drastically change the outcome of the clustering. Above we have selected some types that appear to be beneficial now, however in the future we may add or remove loan metrics. We anticipate to test our clustering by removing a row from the processing, and testing the algorithm to see where it clusters. By repeating this over a large set, we can hopefully determine the most finically potent criteria to include in our clusters.

Peer to Peer Loan Clustering

Johnathan Pasch
Joshua Kruger
Sean Ray

member_id	recoveries	mths_since_recent_bc	
loan_amnt	collection_recovery_fee	mths_since_recent_bc_dlq	
funded_amnt	last_pymnt_d	mths_since_recent_inq	
funded_amnt_inv	last_pymnt_amnt	mths_since_recent_revol_delinq	
term	next_pymnt_d	num_accts_ever_120_pd	
int_rate	last_credit_pull_d	num_actv_bc_tl	
installment	last_fico_range_high	num_actv_rev_tl	
grade	last_fico_range_low	num_bc_sats	
sub_grade	collections_12_mths_ex_med	num_bc_tl	
emp_title	mths_since_last_major_derog	num_il_tl	
emp_length	policy_code	num_op_rev_tl	
home_ownership	application_type	num_rev_accts	
annual_inc	annual_inc_joint	num_rev_tl_bal_gt_0	
verification_status	dti_joint	num_sats	
issue_d	verification_status_joint	num_tl_120dpd_2m	
loan_status	acc_now_delinq	num_tl_30dpd	
pymnt_plan	tot_coll_amt	num_tl_90g_dpd_24m	
url	tot_cur_bal	num_tl_op_past_12m	
desc	open_acc_6m	pct_tl_nvr_dlq	
purpose	open_il_6m	percent_bc_gt_75	
title	open_il_12m	pub_rec_bankruptcies	
zip_code	open_il_24m	tax_liens	
addr_state	mths_since_rcnt_il	tot_hi_cred_lim	
dti	total_bal_il	total_bal_ex_mort	
delinq_2yrs	il_util	total_bc_limit	
earliest_cr_line	open_rv_12m	total_il_high_credit_limit	
fico_range_low	open_rv_24m		
fico_range_high	max_bal_bc		
inq_last_6mths	all_util	Total Number of Loans: 42542	
mths_since_last_delinq	total_rev_hi_lim	Fully Paid Loans: 34350	
mths_since_last_record	inq_fi	Number of available types of	
open_acc	total_cu_tl	criteria: 26	
pub_rec	inq_last_12m		
revol_bal	acc_open_past_24mths		
revol_util	avg_cur_bal		
total_acc	bc_open_to_buy		
initial_list_status	bc_util		
out_prncp	chargeoff_within_12_mths		
out_prncp_inv	delinq_amnt		
total_pymnt	mo_sin_old_il_acct		
total_pymnt_inv	mo_sin_old_rev_tl_op		
total_rec_prncp	mo_sin_rcnt_rev_tl_op		
total_rec_int	mo_sin_rcnt_tl		
total_rec_late_fee	mort_acc		