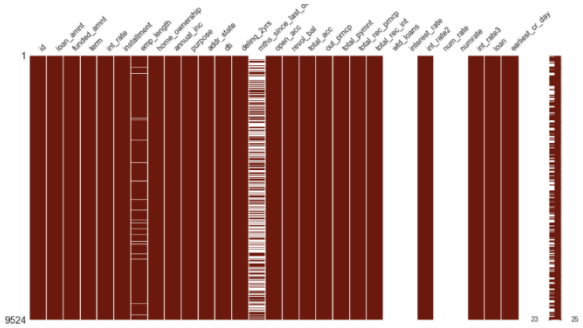
Loan Data Report

Preprocess & Introductory Analysis

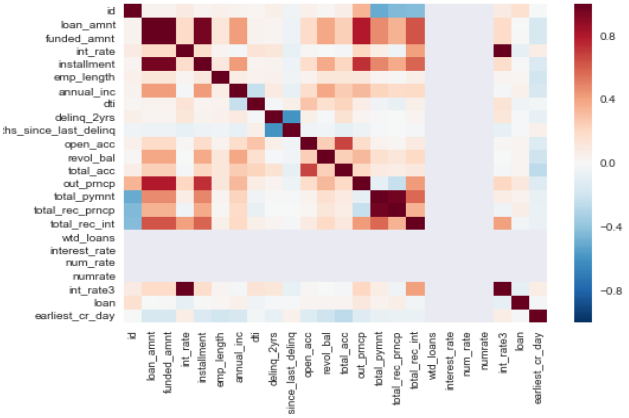
First, the variable "loan" was added to label good as 1 and bad as 0. I encoded "n/a" as missing values. I converted the "earliest_cr_line" to the number of days elapsed since the smallest "earliest_cr_line" value to reduce the number of categorical variables in one hot encoding.

Before doing any analysis, I checked for missing values:



Multiple variables had missing values. The number of good vs bad loans from the missing values of "emp_length" had a very similar distribution as that of the entire dataset so it was reasonable to use the median of "emp_length" to fill the missing values. The other variables had more than 50% of missing data so I dropped them.

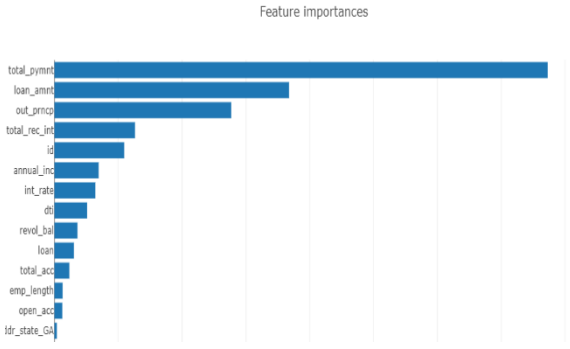
The correlation heat map was shown below:



Next, I checked the correlation between the variables in the data. "loan_amnt", "funded_amnt", "installment" and "total_rec_prncp" were dropped due to high correlation with other variables. The non-float variables were one hot encoded for models to train on.

Preliminary Results

Due to ease of tuning and interpretability, I chose to implement the random forest and logistic regression models. Both models had very good results with 98% accuracy and 0.1 log loss. The precision and recall were also extremely good at around 1 for both. The top features of the random forest model were graphed:



"total_pymnt" and "loan_amnt" were the most important variables. "total_pymnt" made sense logically because people were more likely to pay off loans based on how much they had already paid. "loan_amnt" made sense since smaller loan amounts meant people had easier times paying off the loans on time.

The summary table of the top variables in the logistic model was shown:

	coef	std err	z	P>|z|	[0.025	0.975]
id	5.039e-07	5.64e-08	8.938	0.000	3.93e-07	6.14e-07
loan_amnt	-0.0017	0.000	-8.154	0.000	-0.002	-0.001
int_rate	-0.1177	0.025	-4.662	0.000	-0.167	-0.068
emp_length	-0.0038	0.023	-0.163	0.870	-0.049	0.042
annual_inc	1.549e-06	2.65e-06	0.585	0.558	-3.64e-06	6.74e-06
dti	-0.0262	0.011	-2.314	0.021	-0.048	-0.004
delinq_2yrs	0.1025	0.119	0.859	0.391	-0.131	0.337
open_acc	0.0308	0.023	1.324	0.186	-0.015	0.076
revol_bal	8.362e-06	8.1e-06	1.032	0.302	-7.51e-06	2.42e-05
total_acc	-0.0001	0.010	-0.013	0.990	-0.020	0.020
out_prncp	0.0016	0.000	7.659	0.000	0.001	0.002
total_pymnt	0.0020	0.000	9.226	0.000	0.002	0.002
total_rec_int	-0.0018	0.000	-7.115	0.000	-0.002	-0.001
earliest_cr_day	6.508e-06	2.93e-05	0.222	0.824	-5.1e-05	6.4e-05
term_ 36 months	9.9384	2.32e+04	0.000	1.000	-4.55e+04	4.55e+04
term_ 60 months	10.2861	2.32e+04	0.000	1.000	-4.55e+04	4.55e+04
home_ownership_MORTGAGE	-4.1575	2.78e+04	-0.000	1.000	-5.45e+04	5.44e+04
home_ownership_NONE	16.7516	1.13e+05	0.000	1.000	-2.22e+05	2.22e+05
home_ownership_OTHER	16.2012	1.13e+05	0.000	1.000	-2.22e+05	2.22e+05
home_ownership_OWN	-4.2995	2.78e+04	-0.000	1.000	-5.45e+04	5.44e+04
home_ownership_RENT	-4.2712	2.78e+04	-0.000	1.000	-5.45e+04	5.44e+04
purpose_car	-0.0642	3681.277	-1.74e-05	1.000	-7215.234	7215.106
purpose_credit_card	-0.6003	3681.277	-0.000	1.000	-7215.770	7214.569
purpose_debt_consolidation	-0.6467	3681.277	-0.000	1.000	-7215.816	7214.523
purpose_home_improvement	-0.9960	3681.277	-0.000	1.000	-7216.166	7214.174
purpose_house	-1.3175	3681.277	-0.000	1.000	-7216.487	7213.852
purpose_major_purchase	-1.1215	3681.277	-0.000	1.000	-7216.291	7214.048
purpose_medical	-1.3389	3681.277	-0.000	1.000	-7216.509	7213.831
purpose_moving	0.1622	3681.277	4.41e-05	1.000	-7215.008	7215.332
purpose_other	-0.5661	3681.277	-0.000	1.000	-7215.736	7214.603

The most important features of the logistic regression had p-values less than 0.05. These were the same features as that of the random forest.

Ultimately, I chose to move forward with the logistic regression due to slightly better accuracy (97.7% vs 97.6%) and better interpretability than the random forest model.

Variable Selection

The summary table showed a lot of the one hot encoded variables with p-values close to 1. This meant that I failed to reject the null hypothesis that the coefficients of the associated variables were 0, hence I dropped them. After refitting with fewer variables, I still had variables with p-values greater than 0.05. I wrote a stepwise regression function to drop the variable with the highest p-value and then refitting the model until all variables had p-values less than 0.05. When this occurred, all variables were important in the model.

Final Model & Results

The final model had only 9 variables compared to the original model with more than 80 variables. The results of the final model improved from the original model in accuracy by 0.4% (98%), log loss by 0.002 (0.098) and precision by 0.05% (97.7%).