Geodemographic Segmentation Model

- Eliminated data with no employment length (12068 -> 11064)
- Separated data into two sets
  - ~90%, 10000 of the data to training set
  - ~10%, 1064 of the data to test set

First Model

Model 1: Logit, using observations 1-10000
Dependent variable: loan_status
Standard errors based on Hessian
Omitted due to exact collinearity: term

	coefficient	std. error	z	p-value	
const	-0.0211438	1.11334	-0.01899	0.9848	
loan_amnt	-0.000464050	5.54288e-05	-8.372	5.66e-017	***
installment	0.0138779	0.00158097	8.778	1.66e-018	***
annual_inc	1.00262e-07	5.03997e-07	0.1989	0.8423	
is_inc_v	0.123036	0.0932941	1.319	0.1872	
fico	-0.00317429	0.00158424	-2.004	0.0451	**
dti	0.0154649	0.00344610	4.488	7.20e-06	***
revol_util	-0.00286410	0.00142199	-2.014	0.0440	**
mort_acc	-0.0593558	0.0198963	-2.983	0.0029	***
open_acc_6m	0.0205359	0.0233276	0.8803	0.3787	
inq_last_6mths	0.0501020	0.0336827	1.487	0.1369	
inq_last_12m	-0.00114028	0.0116188	-0.09814	0.9218	
mths_sin_rcnt_bc	-0.00425270	0.00134427	-3.164	0.0016	***
mths_sin_rcnt_inq	-0.00614171	0.00700472	-0.8768	0.3806	
Demp_length_1	-0.124118	0.0798597	-1.554	0.1201	
Demp_length_2	-0.0591435	0.0927686	-0.6375	0.5238	
Demp_length_3	-0.139350	0.0832664	-1.674	0.0942	*

19	Demp_length_1	dummy for emp_length = '0~2 years'
20	Demp_length_2	dummy for emp_length = '10+ years'
21	Demp_length_3	dummy for emp_length = '3~5 years'
22	Demp_length_4	dummy for emp_length = '6~9 years'

Mean dependent var	0.128100	S.D. dependent var	0.334218
McFadden R-squared	0.023895	Adjusted R-squared	0.019454
Log-likelihood	-3736.127	Akaike criterion	7506.253
Schwarz criterion	7628.829	Hannan-Quinn	7547.744

Number of cases 'correctly predicted' = 8716 (87.2%)

Multicollinearity

Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

loan_amnt	340.284
installment	338.236
annual_inc	1.226
is_inc_v	1.022
fico	1.173
dti	1.079
revol_util	1.251
mort_acc	1.170
open_acc_6m	1.127
inq_last_6mths	1.460
inq_last_12m	1.147
mths_sin_rcnt_bc	1.060
mths_sin_rcnt_inq	1.390
Demp_length_1	1.515
Demp_length_2	1.305
Demp_length_3	1.400

Better Adjusted R-squared

Model 2: Logit, using observations 1-10000
Dependent variable: loan_status
Standard errors based on Hessian
Omitted due to exact collinearity: term

	coefficient	std. error	z	p-value	
const	0.609569	1.10830	0.5500	0.5823	
annual_inc	-2.89108e-07	6.12638e-07	-0.4719	0.6370	
is_inc_v	0.157890	0.0929757	1.698	0.0895	*
fico	-0.00419564	0.00157801	-2.659	0.0078	***
dti	0.0165491	0.00346104	4.782	1.74e-06	***
revol_util	-0.00317086	0.00142453	-2.226	0.0260	**
mort_acc	-0.0647932	0.0200651	-3.229	0.0012	***
open_acc_6m	0.0250311	0.0231991	1.079	0.2806	
inq_last_6mths	0.0761927	0.0334084	2.281	0.0226	**
inq_last_12m	0.00161076	0.0115115	0.1399	0.8887	
mths_sin_rcnt_bc	-0.00472420	0.00134881	-3.502	0.0005	***
mths_sin_rcnt_inq	-0.00853195	0.00697395	-1.223	0.2212	
Demp_length_1	-0.131088	0.0794785	-1.649	0.0991	*
Demp_length_2	-0.0579637	0.0923432	-0.6277	0.5302	
Demp_length_3	-0.141451	0.0828652	-1.707	0.0878	*
loan_amnt	1.91956e-05	3.70348e-06	5.183	2.18e-07	***

Mean dependent var	0.128100	S.D. dependent var	0.334218
McFadden R-squared	0.014005	Adjusted R-squared	0.009825
Log-likelihood	-3773.982	Akaike criterion	7579.963
Schwarz criterion	7695.329	Hannan-Quinn	7619.014

Number of cases 'correctly predicted' = 8719 (87.2%)

Model 3: Logit, using observations 1-10000
Dependent variable: loan_status
Standard errors based on Hessian
Omitted due to exact collinearity: term

	coefficient	std. error	z	p-value	
const	0.651080	1.10820	0.5875	0.5569	
annual_inc	-3.98318e-07	6.38407e-07	-0.6239	0.5327	
is_inc_v	0.155966	0.0929962	1.677	0.0935	*
fico	-0.00427206	0.00157757	-2.708	0.0068	***
dti	0.0165064	0.00346841	4.759	1.94e-06	***
revol_util	-0.00322656	0.00142575	-2.263	0.0236	**
mort_acc	-0.0659072	0.0200990	-3.279	0.0010	***
open_acc_6m	0.0250291	0.0232014	1.079	0.2807	
inq_last_6mths	0.0759191	0.0334133	2.272	0.0231	**
inq_last_12m	0.00151212	0.0115200	0.1313	0.8956	
mths_sin_rcnt_bc	-0.00469597	0.00134873	-3.482	0.0005	***
mths_sin_rcnt_inq	-0.00844015	0.00697553	-1.210	0.2263	
Demp_length_1	-0.130927	0.0795015	-1.647	0.0996	*
Demp_length_2	-0.0578500	0.0923734	-0.6263	0.5311	
Demp_length_3	-0.141857	0.0828915	-1.711	0.0870	*
installment	0.000612631	0.000106874	5.732	9.91e-09	***

Mean dependent var	0.128100	S.D. dependent var	0.334218
McFadden R-squared	0.014804	Adjusted R-squared	0.010624
Log-likelihood	-3770.924	Akaike criterion	7573.848
Schwarz criterion	7689.213	Hannan-Quinn	7612.898

Number of cases 'correctly predicted' = 8719 (87.2%)

Backward Elimination Method

```
Model 3: Logit, using observations 1-10000
Dependent variable: loan_status
Standard errors based on Hessian
Omitted due to exact collinearity: term

                     coefficient   std. error      z     p-value
  ---------------------------------------------------------------
  const               0.651080     1.10820       0.5875  0.5569
  annual_inc         -3.98318e-07  6.38407e-07  -0.6239  0.5327
  is_inc_v            0.155966     0.0929962     1.677   0.0935   *
  fico               -0.00427206   0.00157757   -2.708   0.0068   ***
  dti                 0.0165064    0.00346841    4.759   1.94e-06 ***
  revol_util         -0.00322656   0.00142575   -2.263   0.0236   **
  mort_acc           -0.0659072    0.0200990    -3.279   0.0010   ***
  open_acc_6m         0.0250291    0.0232014     1.079   0.2807
  inq_last_6mths      0.0759191    0.0334133     2.272   0.0231   **
  inq_last_12m        0.00151212   0.0115200     0.1313  0.8956
  mths_sin_rcnt_bc   -0.00469597   0.00134873   -3.482   0.0005   ***
  mths_sin_rcnt_inq  -0.00844015   0.00697553   -1.210   0.2263
  Demp_length_1      -0.130927     0.0795015    -1.647   0.0996   *
  Demp_length_2      -0.0578500    0.0923734    -0.6263  0.5311
  Demp_length_3      -0.141857     0.0828915    -1.711   0.0870   *
  installment         0.000612631  0.000106874   5.732   9.91e-09 ***

Mean dependent var   0.128100   S.D. dependent var   0.334218
McFadden R-squared   0.014804   Adjusted R-squared   0.010624
Log-likelihood      -3770.924   Akaike criterion     7573.848
Schwarz criterion    7689.213   Hannan-Quinn         7612.898

Number of cases 'correctly predicted' = 8719 (87.2%)
```

P-value and Adjusted R-squared

Model 10: Logit, using observations 1-10000
Dependent variable: loan_status
Standard errors based on Hessian
Omitted due to exact collinearity: term

	coefficient	std. error	z	p-value	
const	0.635953	1.10329	0.5764	0.5643	
is_inc_v	0.164148	0.0928038	1.769	0.0769	*
fico	-0.00440848	0.00157427	-2.800	0.0051	***
dti	0.0167950	0.00336931	4.985	6.21e-07	***
revol_util	-0.00355914	0.00140842	-2.527	0.0115	**
mort_acc	-0.0722224	0.0191994	-3.762	0.0002	***
inq_last_6mths	0.0985863	0.0284546	3.465	0.0005	***
mths_sin_rcnt_bc	-0.00482948	0.00134486	-3.591	0.0003	***
installment	0.000583500	9.88207e-05	5.905	3.53e-09	***

Mean dependent var	0.128100	S.D. dependent var	0.334218
McFadden R-squared	0.013861	Adjusted R-squared	0.011510
Log-likelihood	-3774.534	Akaike criterion	7567.068
Schwarz criterion	7631.961	Hannan-Quinn	7589.034

Number of cases 'correctly predicted' = 8719 (87.2%)

Model 11: Logit, using observations 1-10000
Dependent variable: loan_status
Standard errors based on Hessian
Omitted due to exact collinearity: term

	coefficient	std. error	z	p-value	
const	0.869321	1.09498	0.7939	0.4272	
fico	-0.00453688	0.00157195	-2.886	0.0039	***
dti	0.0165583	0.00336296	4.924	8.49e-07	***
revol_util	-0.00345150	0.00140762	-2.452	0.0142	**
mort_acc	-0.0746504	0.0191625	-3.896	9.79e-05	***
inq_last_6mths	0.0971651	0.0284287	3.418	0.0006	***
mths_sin_rcnt_bc	-0.00490890	0.00134351	-3.654	0.0003	***
installment	0.000588206	9.89616e-05	5.944	2.79e-09	***

Mean dependent var	0.128100	S.D. dependent var	0.334218
McFadden R-squared	0.013440	Adjusted R-squared	0.011350
Log-likelihood	-3776.146	Akaike criterion	7568.293
Schwarz criterion	7625.976	Hannan-Quinn	7587.818

Number of cases 'correctly predicted' = 8719 (87.2%)

Transformation

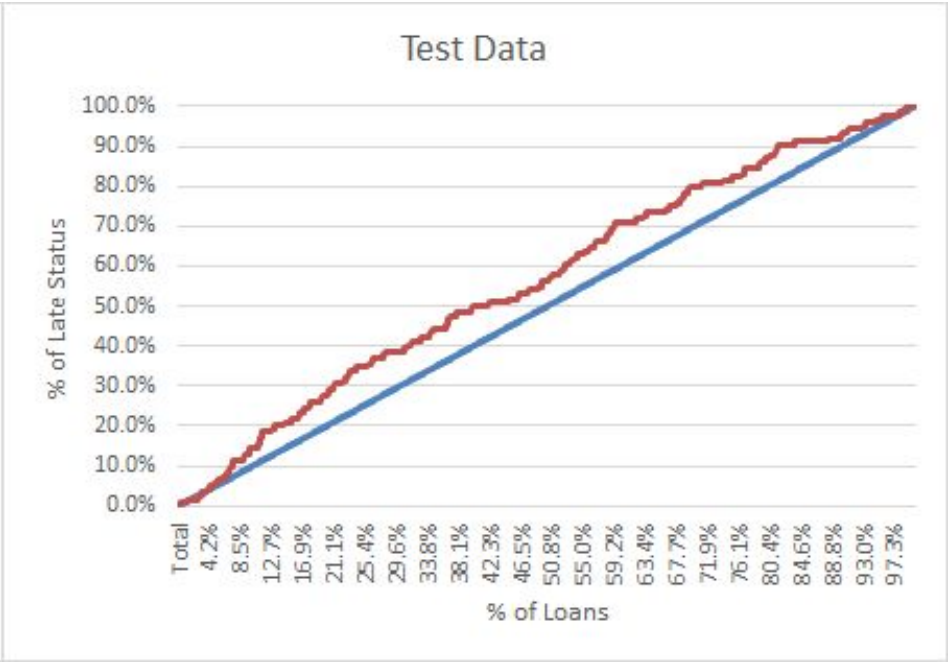
```
Model 12: Logit, using observations 1-10000
Dependent variable: loan_status
Standard errors based on Hessian
Omitted due to exact collinearity: term
```

	coefficient	std. error	z	p-value	
const	-0.868259	1.10124	-0.7884	0.4304	
fico	-0.00421098	0.00156732	-2.687	0.0072	***
dti	0.0163193	0.00336208	4.854	1.21e-06	***
revol_util	-0.00345795	0.00140856	-2.455	0.0141	**
mort_acc	-0.0712173	0.0190445	-3.740	0.0002	***
inq_last_6mths	0.0977986	0.0284453	3.438	0.0006	***
mths_sin_rcnt_bc	-0.00463935	0.00134444	-3.451	0.0006	***
is_inc_v	0.182506	0.0928474	1.966	0.0493	**
log_installment	0.636496	0.0950349	6.697	2.12e-011	***

Mean dependent var	0.128100	S.D. dependent var	0.334218
McFadden R-squared	0.015498	Adjusted R-squared	0.013147
Log-likelihood	-3768.268	Akaike criterion	7554.536
Schwarz criterion	7619.429	Hannan-Quinn	7576.502

```
Number of cases 'correctly predicted' = 8719 (87.2%)
```

Cumulative Accuracy Profile (CAP)



At 50% of loans, 61% of late status can be predicted

At 50% of loans, 57% of late status can be predicted

Odds Ratio

	coefficient	p-value	Odds-ratio
const	-0.868259	0.4304	
fico	-0.00421098	0.0072	0.995797874
dti	0.0163193	1.21E-06	1.016453187
revol_util	-0.00345795	0.0141	0.996548022
mort_acc	-0.0712173	0.0002	0.931259507
inq_last_6mths	0.0977986	0.0006	1.102740671
mths_sin_rcnt_bc	-0.00463935	0.0006	0.995371395
is_inc_v	0.182506	0.0493	1.200221352
log_installment	0.636496	2.12E-11	1.889847239