

DA Group Project

Thangjam Aditya, Semsang D. Bomzon and Naman Dubey

1. Problem Statement

Auto-insurance companies are likely to make policy holders pay insurance irrespective of the driving track records. This increases cost of insurance of good drivers and reduces for bad drivers. XYZ Insurance has tried intervening in this regard with a model to predict if a policy holder will file a claim in the next year. However, the company is not satisfied with the precision and recall metrics of the previous model. Improved precision would mean lesser good drivers requiring to pay while improved recall would mean more bad drivers requiring to pay.

2. Objective

To build a classification model that predicts if a policy holder will file a claim in the next year.

3. Data and Methods

3.1 Data

The dataset contains 5,95,212 observations and 58 useful variables- one is outcome and the remaining 57 are potential predictors. The outcome variable, henceforth referred to as target, is binary- 0, if a policy holder has not filed a claim and 1, if the policy holder has filed a claim.

The dataset is, however, incomplete. While target has zero missing data, 13 of the potential predictors were found to have missing data. The prevalence of missing data among the 57 predictors is given in Figure 1. Also, as seen in Figure 2, 2 predictors have over 40 per cent missing data.

		With Missing Data	Without Missing Data	Total=57
57 Potential Predictors	31 Unordered Categorical	9	22	
	16 Ordered Categorical	1	15	
	10 Continuous	3	7	
		13	44	

Figure 1 Prevalence of missing data in given dataset

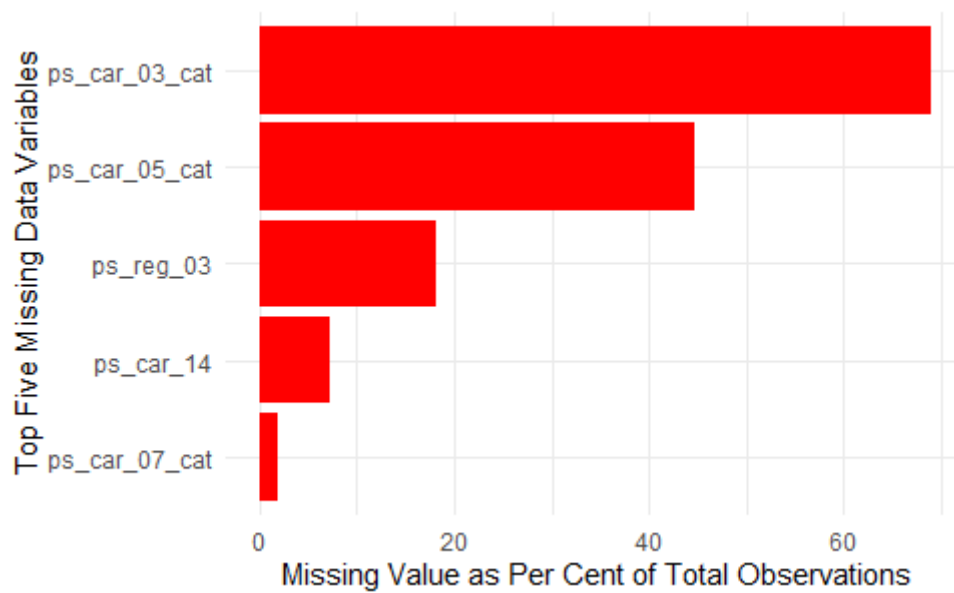


Figure 2 Top five missing data variables

The data is also imbalanced- 96 per cent are non-claimants and 4 per cent are claimants. The methods for handling these two issues as well as the modelling options are discussed in the next subsection.

3.2 Methods

Handling Missing Data

Three cases of handling missing data are considered, namely-

- Case A: Direct list-wise deletion (LWD); this leads to 79 per cent loss of observations
- Case B: Improvised list wise deletion, where LWD was performed after removing “ps-car-03-cat” and “ps-car-05-cat”; this leads to 25 per cent loss of observations
- Case C: Multiple imputation of predictors having missing data except “ps-car-03-cat” and “ps-car-05-cat” followed by LWD; this leads to 34 per cent loss of observations

Dimension Reduction and Feature Selection

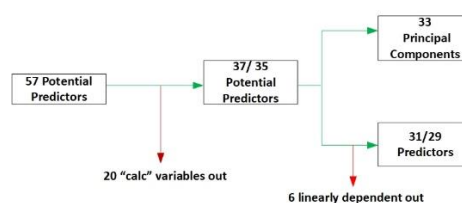


Figure 3 Screening and aggregation of predictors¹²

¹ The six excluded linearly dependent variables are "ps_ind_10_bin", "ps_ind_11_bin", "ps_ind_12_bin", "ps_ind_13_bin", "ps_ind_14", "ps_ind_09_bin"

² Only for Case C, two more variables - “ps-car-03-cat” and “ps-car-05-cat”- are also removed. The reduced numbers are indicated after the slash symbol inside the boxes in the figure.

The variables identified in the first method are excluded from the whole analysis. The linearly dependent predictors from the third method are excluded from models that use original predictors.

Classification Model Choices

Logistic regression, elastic-net regression, gradient boosting and feedforward ANN were selected. Logistic regression is intended to serve as a baseline. The other three were selected because the corresponding R packages enables

- In built model regularization to avoid overfitting
- Faster computational speed for large data (which is the case in hand)

KNN takes a longer amount of time to predict outcomes in large datasets and its value will depend on the size of the class. The predicted outcome will be biased towards the larger number of variables in the dataset. Naïve Bayes assumes that all features are conditionally independent, which may not be the case in real time data.

Handling Class Imbalance

Synthetic random oversampling of minority class (claimants) was done for building training data for logistic regression and elastic-net regression. Boosting and ANN have inbuilt features for dealing with class imbalance.

Key Selected Performance Metrics

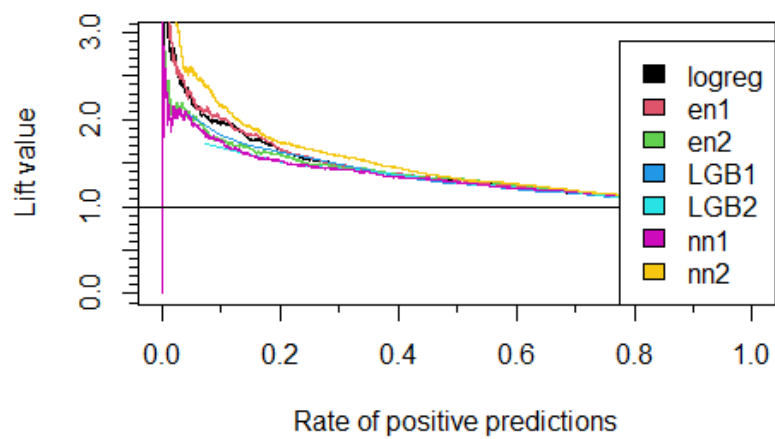
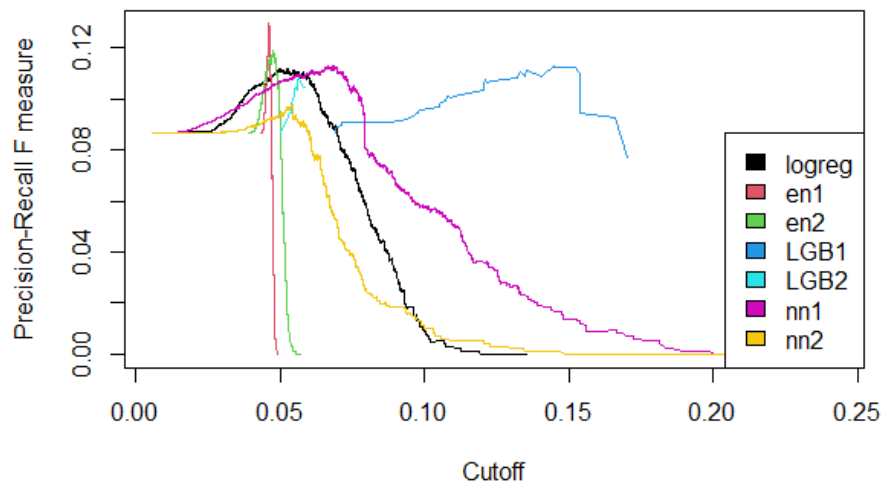
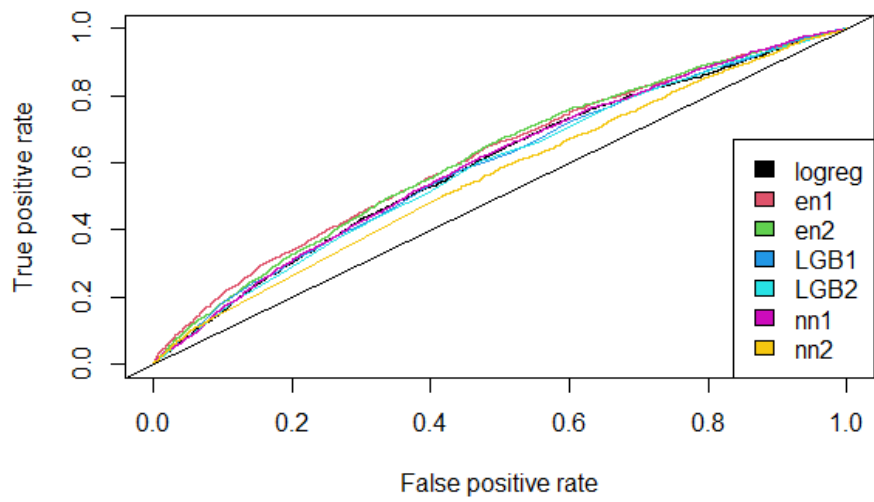
		Actual	
		0 (Claim)	1 (Not Claim)
Predicted	0	Increase True Negatives	Reduce False Negatives
	1	Reduce False Positives	Increase True Positives

Model Selection Criteria

Metric	Priority	Interpretation	Conflicts
Binary Log Loss	(High Priority)	Lower is better	
Gini=2*Area Under ROC curve-1	(High Priority)	Nearer to 1, better it is	
Lift	(High Priority)	Higher than 1, the model is more than “no-model”	
Precision	(Medium Priority)	Higher it is, the better it is	With recall
Recall	(Medium Priority)	Higher it is, better it is	With precision
Accuracy	(Low Priority)	Higher it is, better it is	With the above two

4. Findings

4.1 Case A

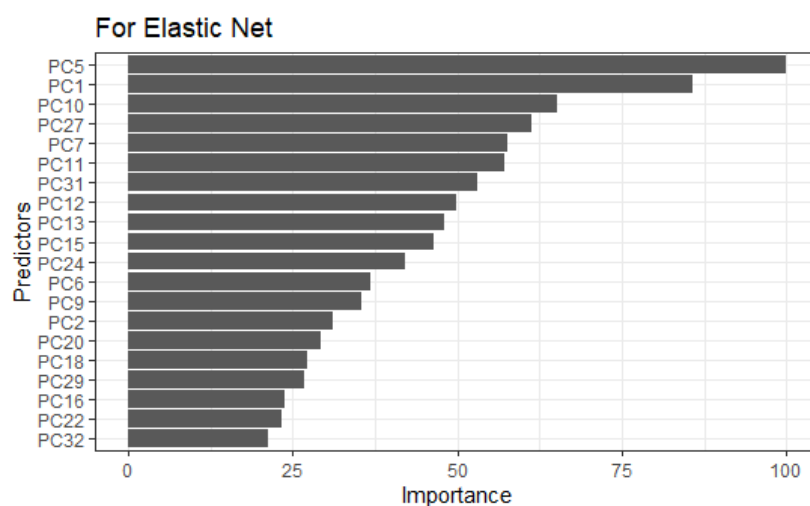


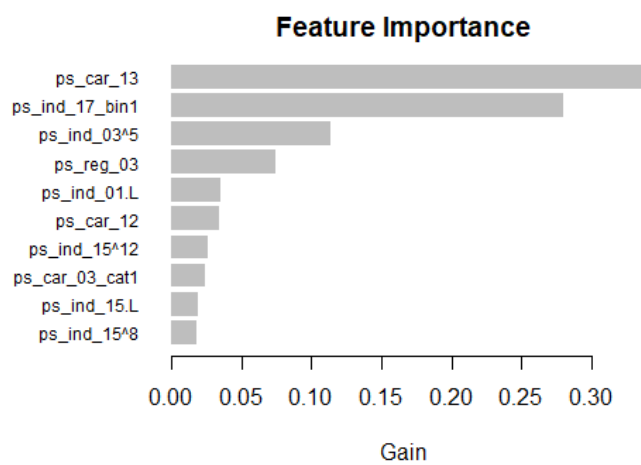
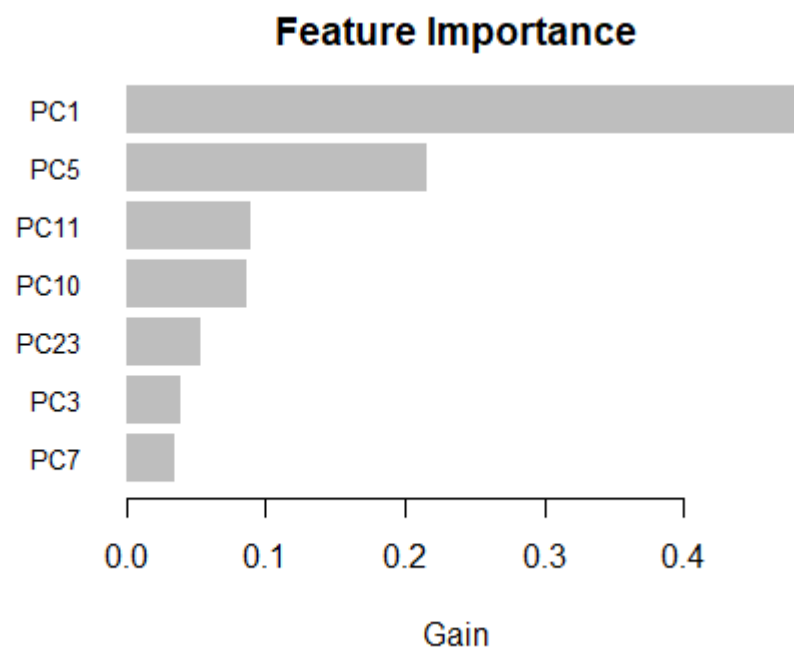
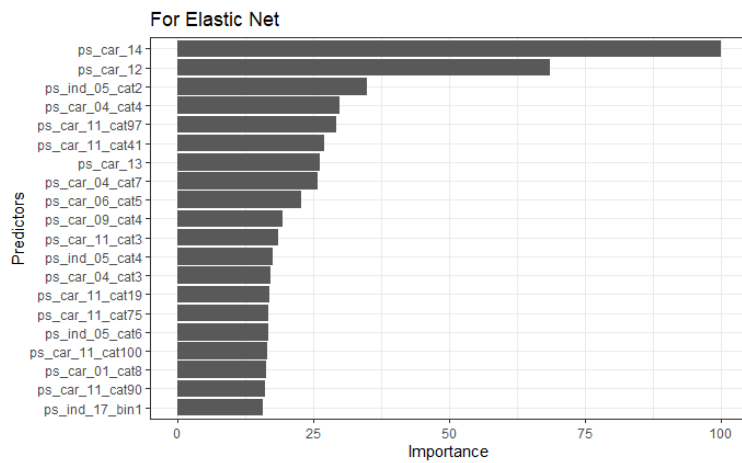
Model Performance

Using PC	Logistic regression		Elastic-net regression		Gradient Boosting	ANN
	UB	B	UB	B		
Training log loss	0.18	0.67	0.18	0.67	0.18	0.18
Validation log loss	0.18	0.67	0.18	0.67	0.18	0.18
Gini measure	0.22	0.21	0.23	0.21	0.20	0.21
Selected cut-off	0.060	0.595	0.046	0.592	0.058	0.067
Precision	0.078	0.080	0.074	0.080	0.077	0.069
Recall	0.304	0.304	0.337	0.232	0.250	0.299
Accuracy (%)	80.5	84.3	78.1	84.56	82.86	78.6

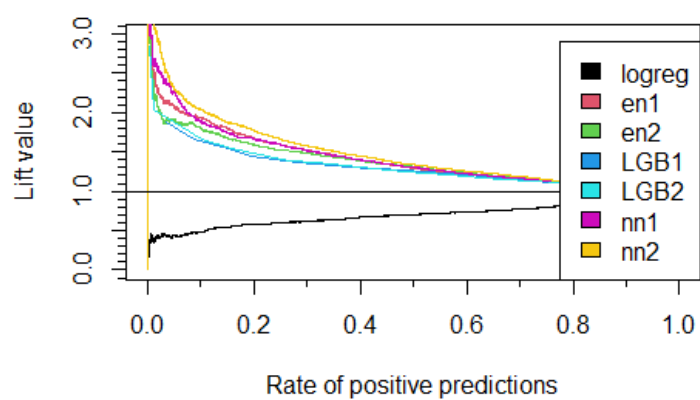
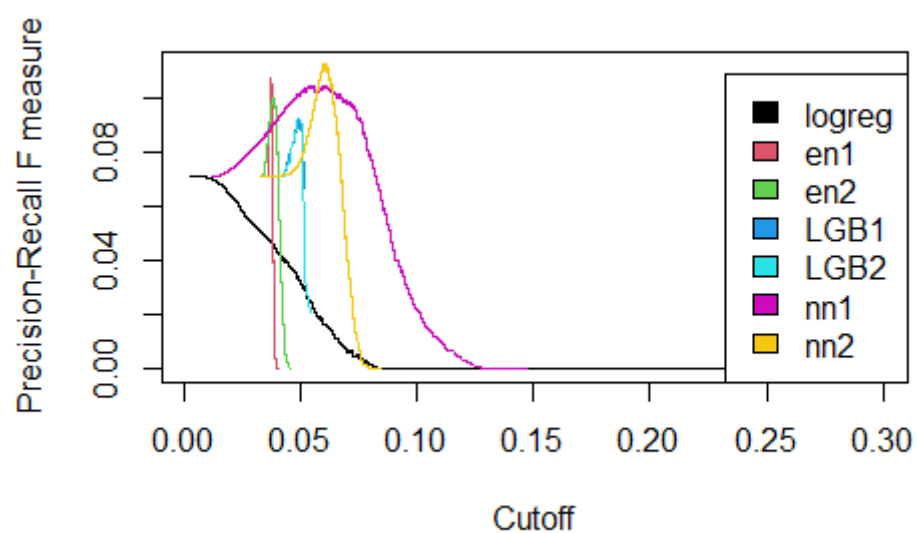
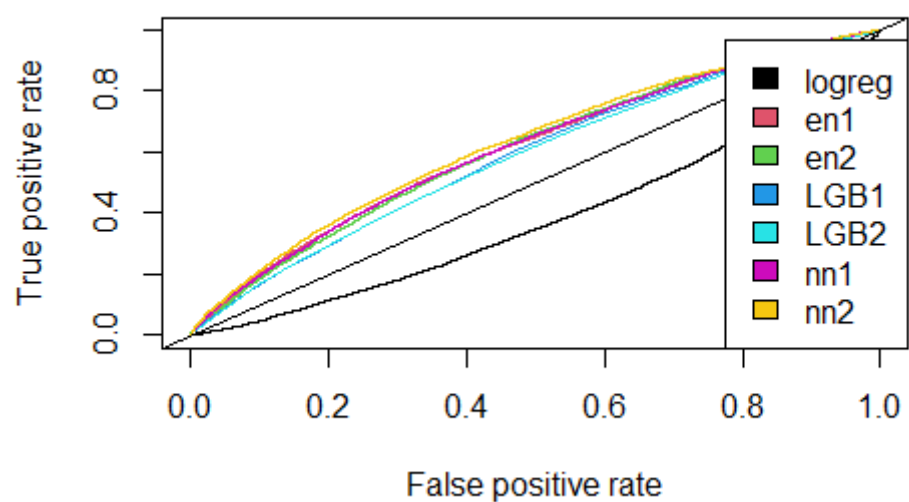
Using original predictors	Logistic regression		Elastic-net regression		Gradient Boosting	ANN
	UB	B	UB	B		
Training log loss	Memory Allocation Problem		0.18	0.66	0.18	0.17
Validation log loss			0.18	0.66	0.18	0.18
Gini measure			0.22	0.26	0.20	0.26
Selected cut-off			0.047	0.63	0.056	0.147
Precision			0.067	0.098	0.060	0.090
Recall			0.37	0.208	0.49	0.230
Accuracy			73.7	87.7	64.8	86

Important Predictors (see the most common predictors with high variable importance scores for each case; and then see the common important predictors among cases; for PCA, see factor loadings to see what makes up each factor)





4.2 Case B

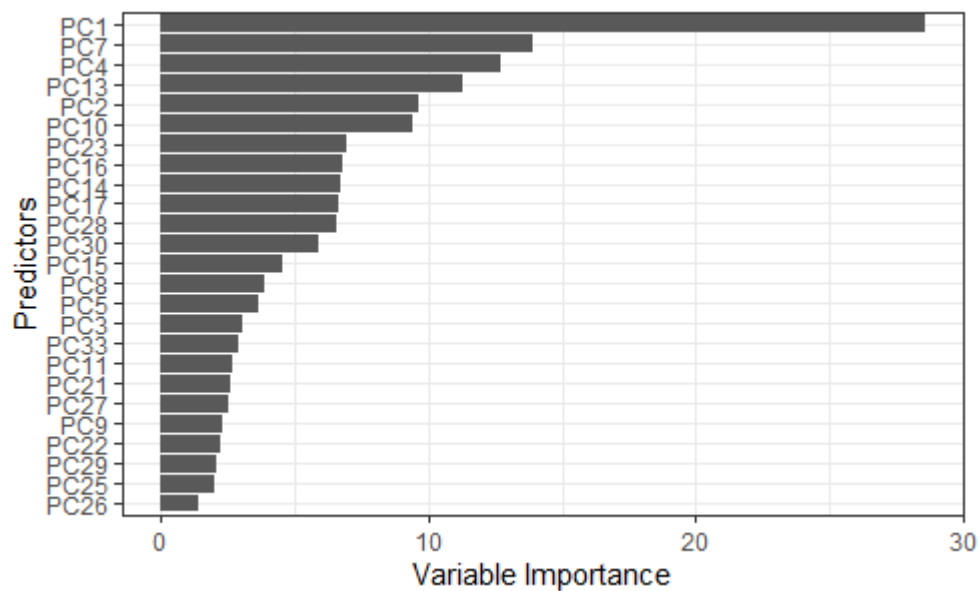


Model Performance

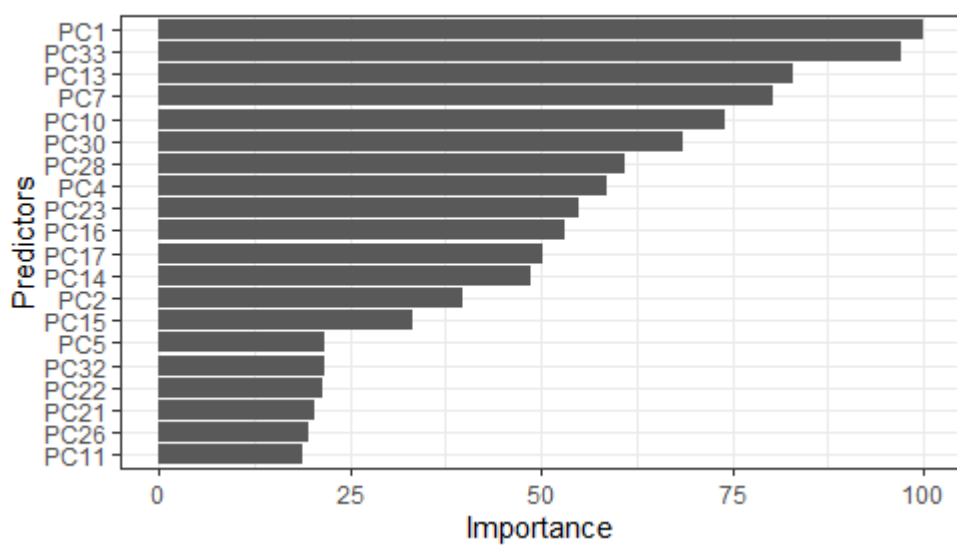
Using PC	Logistic regression		Elastic-net regression		Gradient Boosting	ANN
	UB	B	UB	B		
Training log loss	0.15	Memory Allocation Problem	0.16	0.68	0.16	0.15
Validation log loss	0.15		0.16	0.68	0.16	0.15
Gini measure	-0.226		0.23	0.23	0.18	0.23
Selected cut-off	NC		0.037	0.516	0.049	0.455
Precision	NC		0.063	0.065	0.057	0.067
Recall	NC		0.233	0.28	0.233	0.240
Accuracy	NC		79.8	82.4	82.9	84.7

Using original predictors	Logistic regression		Elastic-net regression		Gradient Boosting	ANN
	UB	B	UB	B		
Training log loss	Memory Allocation Problem		0.16	Memory Allocation Problem	0.16	0.15
Validation log loss			0.16		0.16	0.16
Gini measure			0.22		0.17	0.25
Selected cut-off			0.038		0.049	0.061
Precision			0.056		0.057	0.073
Recall			0.397		0.242	0.245
Accuracy			73.14		82.4	85.6

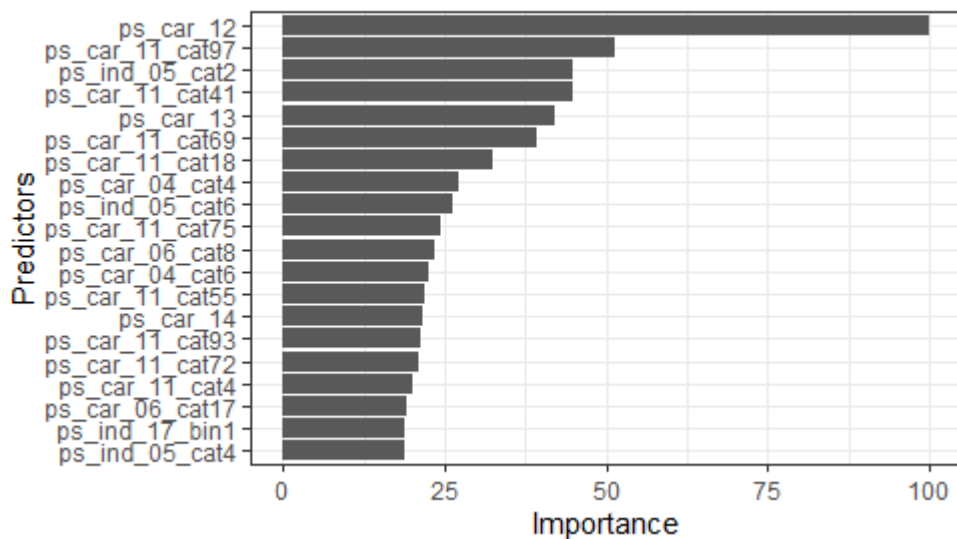
Important Predictors



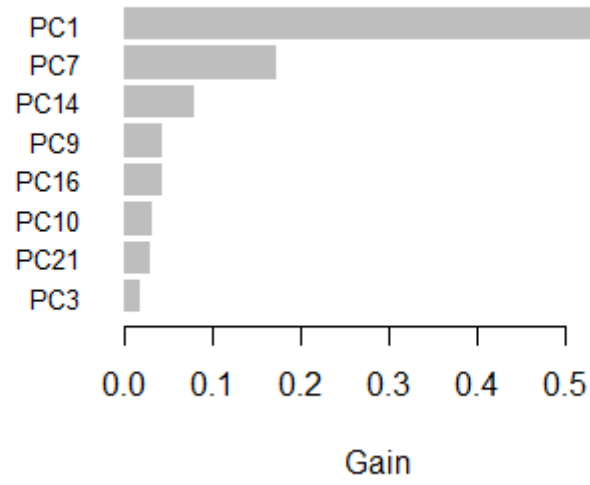
For Elastic Net



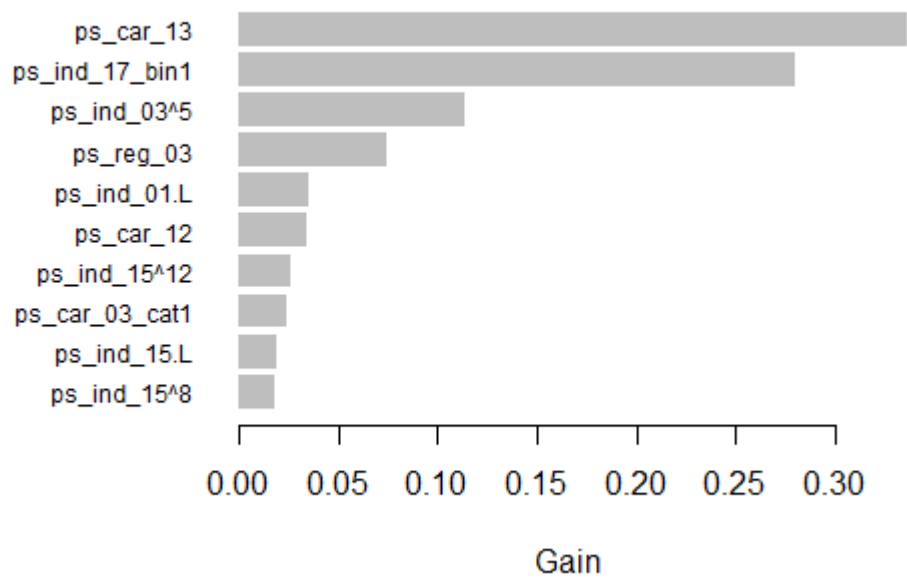
For Elastic Net



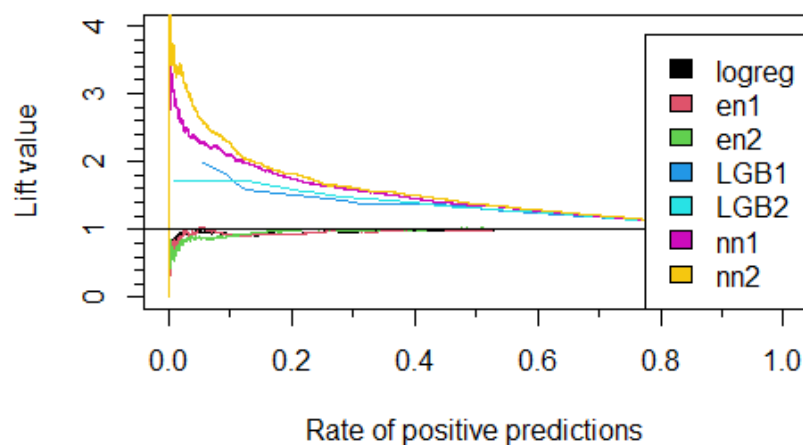
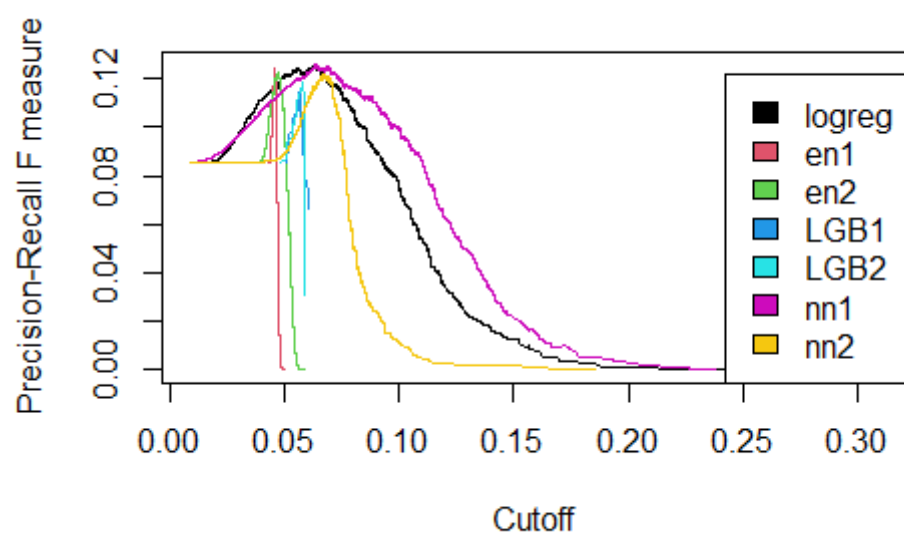
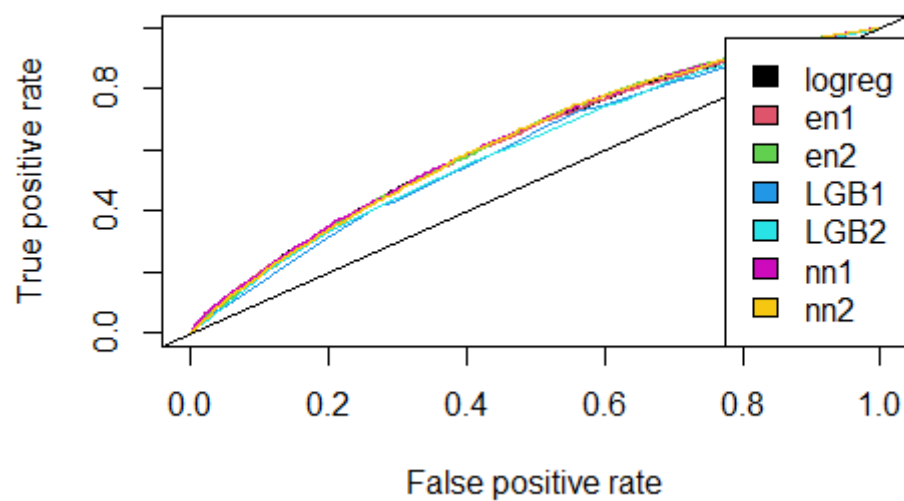
Feature Importance



Feature Importance



4.3 Case C

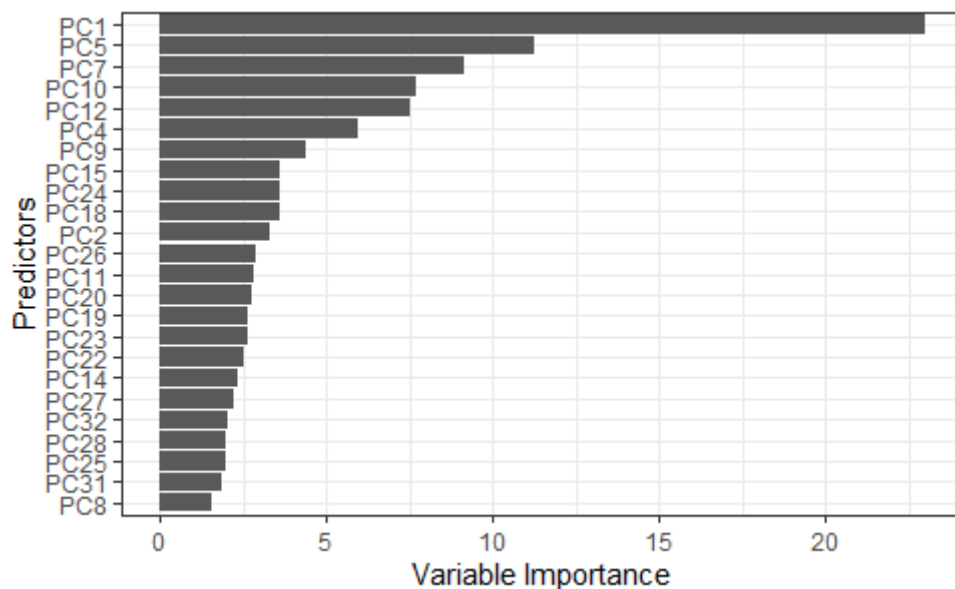


Model Performance

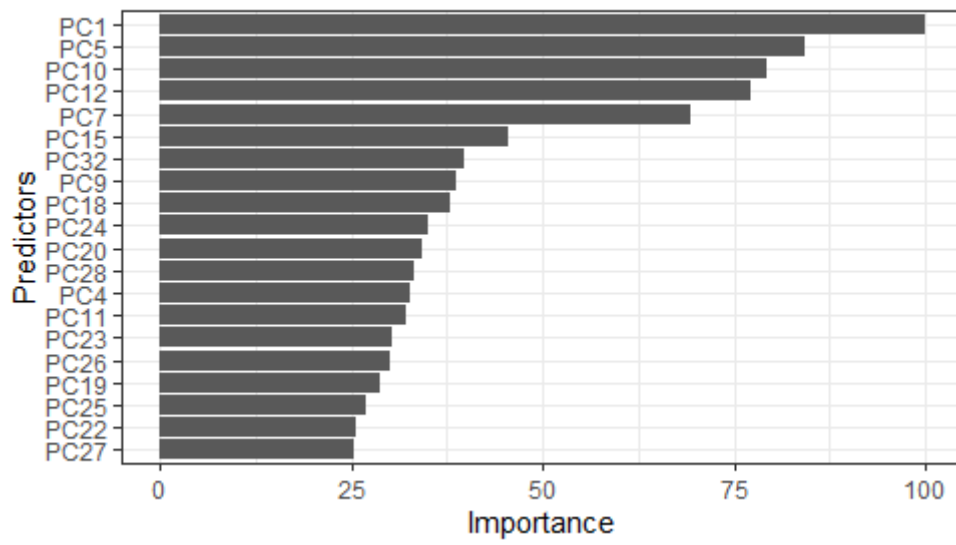
Using PC	Logistic regression		Elastic-net regression		Gradient Boosting	ANN
	UB	B	UB	B		
Training log loss	0.18	Memory Allocation Problem	0.18	Memory Allocation Problem	0.18	0.18
Validation log loss	0.18		0.18		0.18	0.18
Gini measure	0.25		0.25		0.22	0.27
Selected cut-off	0.057		0.045		0.056	0.050
Precision	0.074		0.060		0.060	0.080
Recall	0.343		0.529		0.43	0.366
Accuracy	77.8		63.3		67.9	84.2

Using original predictors	Logistic regression		Elastic-net regression		Gradient Boosting	ANN
	UB	B	UB	B		
Training log loss	Memory Allocation Problem		0.18	Memory Allocation Problem	0.18	0.17
Validation log loss			0.18		0.18	0.18
Gini measure			0.25		0.22	0.28
Selected cut-off			0.046		0.056	0.137
Precision			0.068		0.064	0.10
Recall			0.468		0.470	0.22
Accuracy			69.3		67.2	87.8

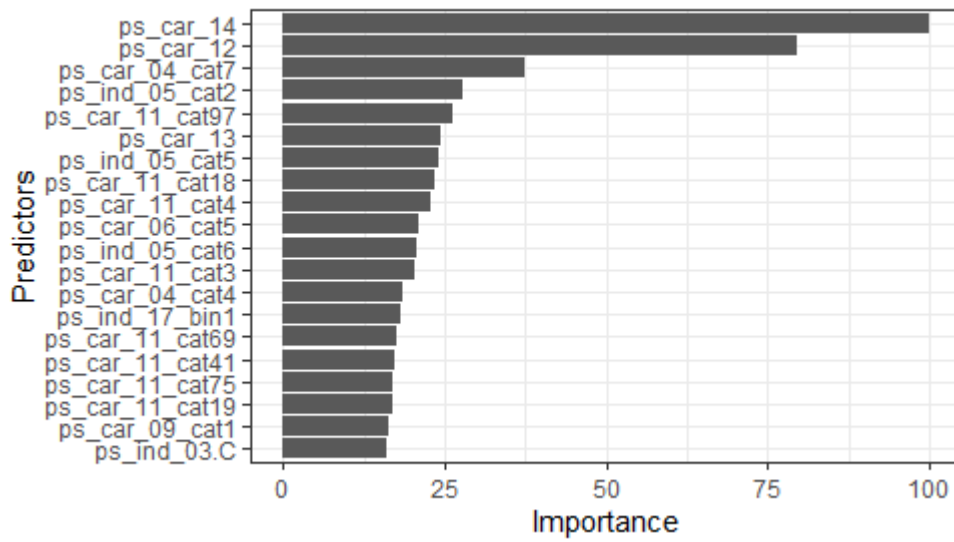
Important Predictors

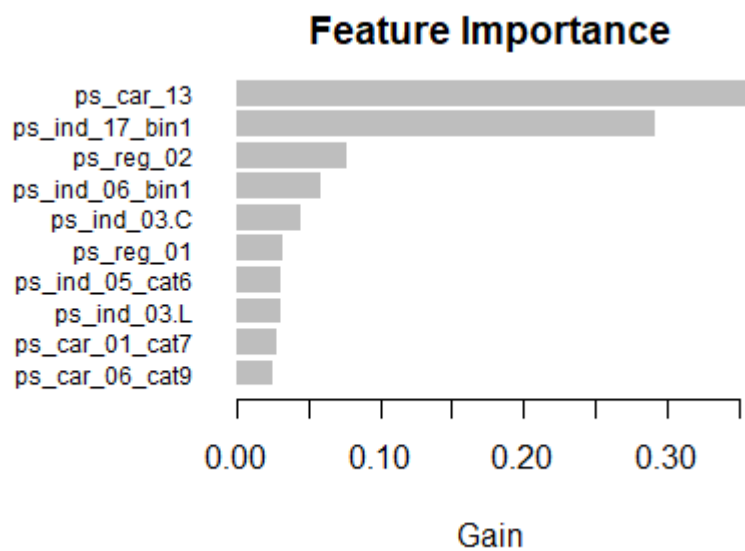
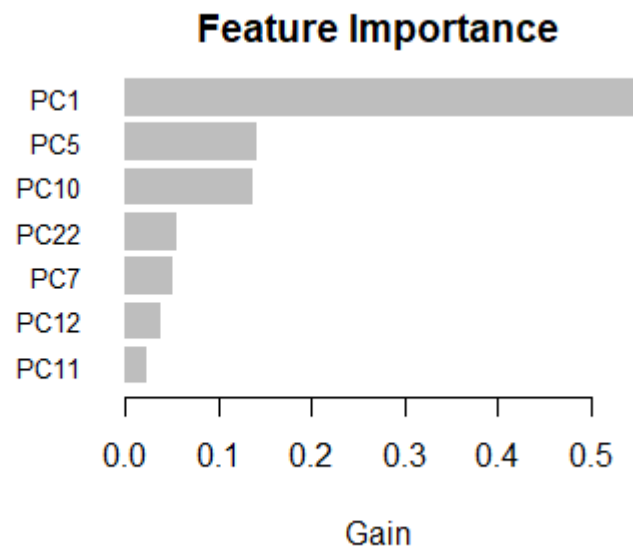


For Elastic Net



For Elastic Net



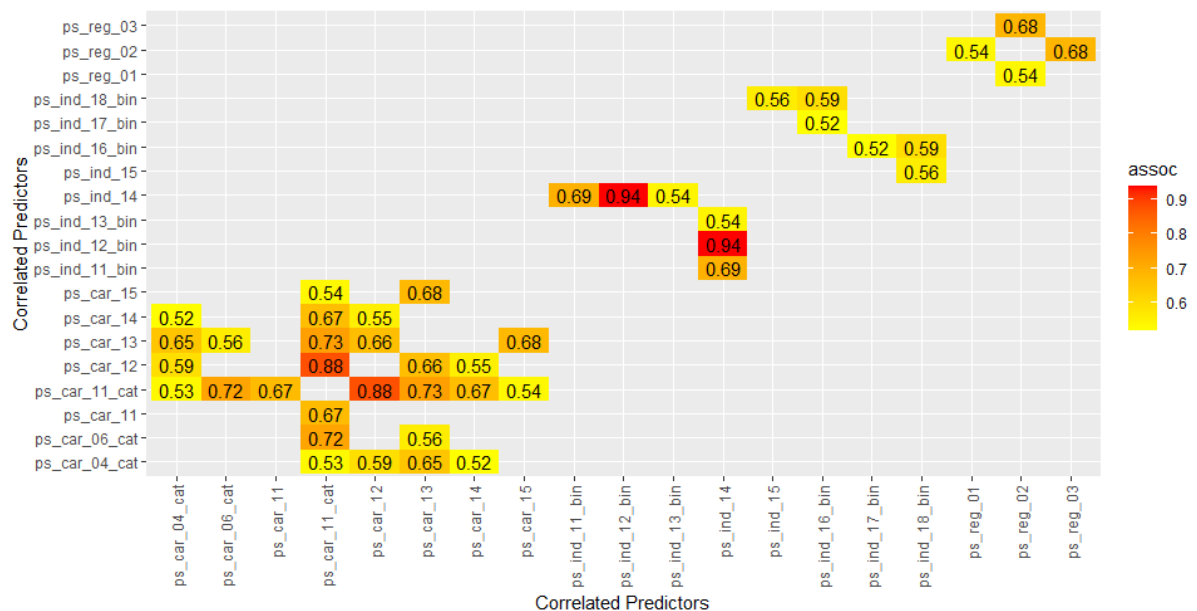


5. Recommendations

6. Limitations

Appendix

Key Package	Use
Caret	Step wise logistic regression Parameter tuning of elastic-net
glmnet	Elastic-net modelling
Light GBM	Gradient Boosting
Keras	ANN



Factor Loadings(DARKER SHADES INDICATE STRONG LOADING IRRESPECTIVE OF SIGN)

Factor Loadings

Var	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24	PC25	PC26	PC27	PC28	PC29	PC30	PC31	PC32	PC33	Loading
ps_car_01_cat	-0.197	0.066	-0.053	0.210	-0.016	0.069	0.096	0.047	0.077	0.105	0.093	0.017	-0.026	-0.032	0.001	-0.067	0.251	0.254	0.173	0.190	-0.337	-0.076	-0.292	-0.506	-0.341	-0.256	0.148	0.000	-0.027	-0.011	0.032	-0.004	-0.028	
ps_car_02_cat	0.330	-0.012	0.011	0.076	-0.029	-0.034	-0.062	-0.110	-0.007	-0.052	-0.060	-0.019	-0.028	-0.019	-0.022	-0.069	-0.031	0.092	0.035	-0.040	-0.024	0.004	0.001	-0.057	0.303	-0.156	0.622	0.385	0.071	-0.261	-0.242	0.086	-0.010	
ps_car_03_cat	-0.087	0.099	0.284	0.221	0.094	0.218	-0.297	-0.034	-0.012	-0.110	0.052	-0.011	-0.065	0.077	0.245	-0.011	-0.092	-0.008	-0.060	-0.138	-0.111	0.195	0.262	-0.059	0.069	-0.291	0.080	0.043	-0.014	0.607	-0.028	0.026	0.005	
ps_car_04_cat	-0.325	-0.035	-0.047	-0.262	0.060	-0.037	0.006	-0.050	-0.068	0.054	-0.067	0.000	0.005	-0.038	-0.091	-0.004	-0.100	0.008	-0.007	0.038	0.025	-0.005	-0.033	-0.060	0.264	-0.143	0.159	0.271	0.267	0.046	0.681	0.106	0.017	
ps_car_05_cat	-0.039	0.106	-0.309	0.053	0.055	-0.040	0.031	0.051	-0.136	-0.203	0.107	0.130	0.059	0.055	0.238	0.151	-0.051	-0.109	-0.009	0.251	-0.678	0.077	0.050	0.162	0.190	0.283	-0.011	0.076	0.102	-0.009	-0.028	-0.022	0.001	
ps_car_06_cat	-0.164	0.015	-0.091	-0.092	0.003	0.023	0.017	0.079	0.087	0.337	0.062	-0.346	-0.155	0.126	0.135	0.423	0.123	-0.072	-0.109	-0.466	-0.143	-0.045	-0.006	0.201	-0.247	0.088	0.274	0.091	-0.036	-0.034	0.011	-0.028	-0.007	
ps_car_07_cat	0.010	0.052	0.039	0.098	0.087	-0.091	0.101	0.336	-0.203	0.240	0.406	0.073	0.160	0.085	0.105	0.263	0.111	0.436	0.142	0.117	0.293	0.179	0.239	-0.053	0.197	0.094	0.002	0.027	0.027	-0.025	0.003	-0.003	0.009	
ps_car_08_cat	0.090	0.026	-0.421	-0.090	-0.038	-0.272	0.096	0.060	-0.025	-0.020	-0.004	0.122	0.087	-0.081	-0.139	-0.121	0.093	-0.014	-0.018	-0.153	0.133	0.046	0.032	-0.091	-0.266	0.092	-0.090	0.435	0.212	0.478	-0.140	-0.023	0.012	
ps_car_09_cat	0.010	0.007	-0.114	0.039	-0.094	-0.023	0.020	0.191	0.071	0.287	0.134	-0.213	-0.668	0.058	0.129	-0.415	-0.257	0.021	-0.003	0.232	0.026	0.042	0.022	0.055	0.054	0.067	-0.078	0.084	0.037	0.015	-0.025	-0.022	0.001	
ps_car_10_cat	-0.018	0.016	0.081	0.056	0.032	-0.006	-0.076	0.020	0.080	-0.084	0.336	0.478	-0.474	-0.211	-0.467	0.282	0.053	-0.113	-0.156	-0.107	-0.051	-0.006	-0.035	0.052	0.003	0.025	-0.031	0.000	-0.023	-0.007	0.014	0.006		
ps_car_11	-0.048	0.024	0.135	0.145	0.011	-0.027	0.113	0.375	-0.051	0.120	0.133	0.170	0.317	-0.100	-0.180	-0.312	-0.496	0.043	-0.023	-0.259	-0.178	-0.119	-0.118	0.204	-0.126	0.044	0.214	-0.083	-0.081	0.037	0.057	0.028	-0.001	
ps_car_11_cat	-0.081	-0.025	-0.066	-0.085	0.001	-0.139	0.064	-0.128	0.179	-0.361	-0.201	0.035	-0.217	0.074	0.034	0.169	-0.338	0.678	0.125	-0.190	-0.042	-0.032	-0.004	0.073	-0.062	0.012	-0.065	-0.062	0.033	0.006	-0.058	-0.013	0.003	
ps_car_12	-0.364	-0.011	-0.105	-0.252	0.048	0.003	0.013	0.011	-0.076	0.083	-0.006	0.020	0.032	-0.025	-0.037	-0.031	-0.043	-0.042	-0.018	0.033	0.013	0.028	0.037	-0.039	0.077	-0.097	0.009	-0.046	-0.075	-0.002	-0.489	0.671	-0.016	
ps_car_13	-0.403	-0.022	0.199	-0.123	0.066	0.089	-0.030	0.002	-0.024	0.012	-0.030	0.025	0.007	-0.003	-0.018	0.012	-0.057	0.032	0.006	0.092	-0.003	-0.003	-0.017	0.022	-0.016	0.089	-0.142	0.077	0.083	-0.098	-0.014	0.059	-0.009	
ps_car_14	-0.349	-0.026	-0.074	-0.269	0.052	-0.004	0.018	0.011	-0.090	0.080	-0.028	0.035	0.051	-0.041	-0.070	-0.030	-0.096	-0.048	-0.031	0.015	0.009	0.022	0.028	-0.088	0.198	-0.194	0.092	-0.025	-0.046	-0.006	-0.363	-0.718	-0.015	
ps_car_15	-0.131	-0.004	0.444	0.131	0.044	0.167	-0.146	-0.039	0.072	0.033	-0.035	0.004	0.053	-0.018	-0.041	-0.025	0.044	0.065	0.035	0.065	-0.010	-0.057	-0.090	0.095	-0.153	0.327	-0.140	0.466	0.348	-0.117	-0.217	-0.076	-0.008	
ps_ind_01	-0.124	0.154	-0.252	0.141	-0.053	0.017	-0.332	0.077	-0.065	-0.112	0.038	0.102	0.042	0.012	0.145	-0.169	-0.080	-0.051	-0.099	-0.333	0.044	0.211	0.276	-0.212	-0.168	-0.126	-0.201	0.096	0.049	-0.526	0.082	-0.009	-0.017	
ps_ind_02_cat	0.017	0.030	0.001	0.231	-0.026	-0.207	0.025	-0.213	0.008	0.180	-0.079	-0.086	0.097	0.089	-0.101	0.399	-0.556	-0.174	-0.184	0.321	0.095	0.164	0.016	-0.245	-0.204	-0.010	0.020	0.076	-0.044	-0.013	-0.019	0.010	0.004	
ps_ind_03	-0.040	0.092	-0.180	0.018	0.053	0.257	-0.177	-0.047	-0.428	-0.257	0.193	-0.054	-0.092	0.062	0.139	0.040	-0.140	-0.041	-0.039	-0.007	0.325	-0.215	-0.358	-0.116	-0.098	0.219	0.235	-0.205	0.234	0.070	-0.038	-0.011	0.011	
ps_ind_04_cat	-0.030	0.036	0.093	0.266	-0.023	-0.483	-0.047	0.165	-0.002	0.001	-0.004	-0.159	-0.003	0.027	0.055	0.104	0.041	-0.117	-0.018	-0.147	-0.041	-0.216	-0.242	0.018	0.264	-0.315	-0.234	-0.156	0.452	-0.014	-0.124	0.030	-0.005	
ps_ind_05_cat	0.000	0.000	0.002	-0.011	-0.005	0.029	-0.074	0.163	-0.117	-0.359	0.188	-0.681	0.051	-0.210	-0.467	0.008	0.011	0.062	-0.008	0.070	-0.119	0.115	0.124	-0.024	-0.011	-0.005	-0.054	0.039	-0.062	-0.004	0.001	0.000	0.003	
ps_ind_06_bin	0.225	-0.033	-0.010	-0.227	0.022	0.386	0.263	0.291	0.205	-0.106	0.048	0.014	0.035	0.054	0.082	0.154	-0.147	-0.091	-0.070	0.025	0.003	0.021	-0.024	-0.085	-0.043	-0.243	-0.115	0.057	0.241	-0.052	-0.048	0.018	-0.008	
ps_ind_07_bin	-0.200	-0.100	0.256	0.065	-0.061	-0.419	0.110	0.109	-0.220	-0.261	-0.105	0.026	-0.149	0.025	0.156	-0.072	0.125	-0.120	-0.024	-0.016	0.062	0.042	0.037	-0.100	-0.080	0.183	0.175	0.071	-0.265	-0.011	0.047	-0.010	0.003	
ps_ind_08_bin	-0.051	0.044	-0.164	-0.034	0.047	-0.131	-0.477	-0.236	0.438	0.148	0.295	-0.060	0.202	-0.013	-0.048	-0.093	-0.020	0.086	0.070	0.031	0.012	-0.074	-0.094	0.000	0.141	0.182	0.035	-0.098	-0.056	0.068	0.011	-0.009	0.026	
ps_ind_09_bin	0.014	0.126	-0.141	0.257	0.000	0.168	0.023	-0.276	-0.465	0.320	-0.249	0.012	-0.073	-0.090	-0.261	-0.013	0.057	0.184	0.050	-0.046	-0.097	-0.002	0.086	0.244	0.008	-0.110	-0.110	-0.063	0.082	0.010	-0.009	-0.002	-0.021	
ps_ind_10_bin	-0.021	-0.132	-0.047	0.053	0.117	0.022	-0.049	0.074	-0.010	-0.056	-0.025	0.044	-0.001	0.868	-0.379	-0.105	0.074	-0.029	-0.018	0.002	-0.050	-0.095	0.084	-0.038	0.014	0.009	0.013	-0.002	0.006	0.001	-0.002	-0.001	0.000	
ps_ind_11_bin	-0.003	-0.301	-0.097	0.115	0.213	0.033	-0.070	0.089	0.017	0.030	-0.084	-0.007	-0.031	-0.216	0.063	0.077	-0.063	-0.033	0.053	0.074	-0.012	-0.647	0.489	-0.157	-0.041	0.046	0.025	0.002	0.018	0.006	0.004	0.004	-0.001	
ps_ind_12_bin	-0.014	-0.419	-0.110	0.132	0.204	0.029	-0.075	0.064	0.001	0.008	-0.052	0.003	0.033	-0.106	0.077	-0.032	0.083	0.163	-0.419	0.038	0.024	0.286	-0.288	0.115	0.020	-0.047	-0.023	0.021	-0.046	-0.008	0.002	0.000	-0.003	
ps_ind_13_bin	-0.012	-0.225	-0.075	0.087	0.181	0.034	-0.073	0.076	0.009	0.001	-0.081	0.012	-0.076	-0.047	-0.042	0.104	-0.116	-0.272	0.789	-0.114	0.040	0.300	-0.121	0.019	-0.009	0.016	0.002	-0.011	0.008	-0.004	-0.006	-0.003	0.002	
ps_ind_14	-0.017	-0.492	-0.141	0.167	0.286	0.044	-0.105	0.107	0.007	0.008	-0.091	0.010	-0.004	-0.031	0.012	0.010	0.025	0.043	-0.112	0.026	0.016	0.062	-0.074	0.035	0.002	-0.015	-0.007	0.014	-0.026	-0.005	0.001	0.000	-0.002	
ps_ind_15	0.040	0.334	-0.082	0.024	0.233	0.038	-0.170	0.214	-0.055	-0.044	-0.104	0.018	-0.037	0.029	0.025	0.151	-0.042	0.004	0.033	0.055	0.096	-0.273	-0.300	0.109	0.113	-0.164	-0.231	0.392	-0.513	0.001	0.031	0.000	0.011	
ps_ind_16_bin	0.065	0.282	0.049	-0.035	0.581	-0.087	0.133	-0.034	0.078	0.024	-0.094	-0.057	-0.052	-0.042	-0.020	-0.117	0.032	-0.020	-0.070	-0.042	-0.001	0.109	0.059	-0.101	-0.031	0.113	0.074	-0.145	0.128	-0.036	-0.004	-0.005	0.006	
ps_ind_17_bin	-0.068	-0.054	-0.080	0.025	-0.522	0.079	-0.310	0.339	0.050	0.008	-0.281	0.084	0.022	-0.017	-0.069	0.141	0.051	0.041	0.012	0.166	0.042	0.017	0.012	0.074	0.001	0.026	0.122	-0.077	0.054	0.073	0.000	-0.004	-0.005	
ps_ind_18_bin	-0.009	-0.333	0.028	0.022	-0.195	0.026	0.177	-0.351	-0.157	-0.034	0.454	-0.026	0.040	0.073	0.107	-0.031	-0.098	-0.025	0.073	-0.142	-0.054	-0.128	-0.051	0.031	0.027	-0.155	-0.200	0.202	-0.126	-0.022	-0.001	-0.001	0.001	
ps_reg_01	-0.125	0.025	-0.090	0.294	-0.143	0.239	0.270	0.011	0.139	0.033	-0.160	-0.083	0.019																					