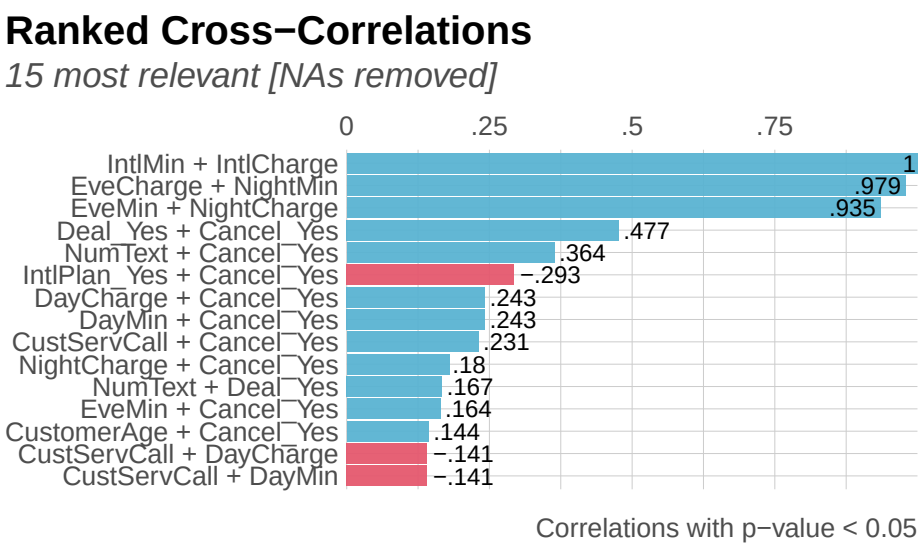
OS 4016 Project II: Cell Phone Plan Cancellations

Doug Andrade

Exploratory Data Analysis and Preparation

The data was reviewed for class, NaNs, correlation, significance, distribution, and balance. I performed feature transformations, one-hot encoding, resampling, and feature reduction (manual and PCA).



In cross-correlation analysis, I identified multicollinearity (`DayMin`, `EveMin`, `EveCharge`, `IntlMin`) and categorical features of $> .05$ p-value relative to `Cancel` (`Married`, `PaymentMethod`). I considered these features for removal in during modeling.

To address non-normal & uniform distributions, the `bestNormalization()` function to apply the best transformations (Ordered Quantile, Square Root, and exponential, logarithmic) on training, validation & test features for optimal predictive power.

I over-sampled the data based on the target variable `Cancel` from `No:554, Yes:255` to `No:554, Yes:533`. I one-hot encoded categorical features, excluding `Cancel`. PCA analysis was evaluated for feature reduction and model simplicity. The 1st component captured little variance in the data (0.1075), resulting in much information lost.

Table 1: Normalized One-Hot Encoded Training Data Stats

	Mean	Median	StDev	Min	Max	Skew	Total.NA
CustomerAge	0.0109	0.0636	1.0565	-3.2304	3.2304	0	0
Married.divorced	0.2328	0.0000	0.4228	0.0000	1.0000	0	0
Married.married	0.5299	1.0000	0.4993	0.0000	1.0000	0	0
Married.single	0.2374	0.0000	0.4257	0.0000	1.0000	0	0
HouseholdSize	0.0015	-0.5187	1.0094	-0.8325	1.7995	0	0
AccountAge	110.7065	110.0000	43.9043	1.0000	249.0000	0	0
PaymentMethod.Automatic	0.7075	1.0000	0.4551	0.0000	1.0000	0	0
PaymentMethod.Check	0.1785	0.0000	0.3831	0.0000	1.0000	0	0
PaymentMethod.Credit	0.1141	0.0000	0.3180	0.0000	1.0000	0	0
LastNewPhone	-0.0084	-0.0387	0.9601	-2.5388	2.3802	0	0
BasePlan.deluxe	0.1150	0.0000	0.3192	0.0000	1.0000	0	0
BasePlan.economy	0.3027	0.0000	0.4596	0.0000	1.0000	0	0
BasePlan.standard	0.5823	1.0000	0.4934	0.0000	1.0000	0	0
IntlPlan.No	0.1757	0.0000	0.3808	0.0000	1.0000	0	0
IntlPlan.Yes	0.8243	1.0000	0.3808	0.0000	1.0000	0	0
Deal.No	0.4563	0.0000	0.4983	0.0000	1.0000	0	0
Deal.Yes	0.5437	1.0000	0.4983	0.0000	1.0000	0	0
CustServCall	0.0831	0.4110	1.0449	-1.6288	2.9324	0	0
NumVmail	-0.0053	-0.5334	1.0340	-0.5334	4.9717	0	0
NumText	0.1389	0.1548	1.0187	-1.3007	2.9833	0	0
NumApps	-0.0187	-0.0651	0.9818	-2.2654	2.2654	0	0

DayMin	262.6169	261.7000	86.7689	37.2000	482.4000	0	0
DayCall	65.3910	66.0000	12.8781	23.0000	97.0000	0	0
DayCharge	42.8575	42.7100	14.1611	6.0600	78.7200	0	0
EveMin	212.8819	215.4000	52.2120	66.1000	364.1000	0	0
EveCall	76.4591	78.0000	14.9110	33.0000	114.0000	0	0
EveCharge	11.3595	11.3700	2.6608	1.2400	22.4000	0	0
NightMin	157.3753	157.6000	37.4535	18.8000	309.3000	0	0
NightCall	91.8160	92.0000	18.2696	34.0000	146.0000	0	0
NightCharge	18.8548	18.5500	4.8561	3.8000	34.3900	0	0
IntlMin	10.9874	11.0000	2.9524	0.0000	20.8000	0	0
IntlCall	-0.0297	-0.0078	0.9792	-1.6833	5.4379	0	0
IntlCharge	3.0808	3.0800	0.8281	0.0000	5.8300	0	0
DayData	0.0292	0.0046	0.9887	-2.7380	2.9020	0	0
EveData	-0.0059	0.0170	0.9710	-3.2304	2.6775	0	0
NightData	-0.0019	-0.0077	0.9813	-2.8106	3.0267	0	0

Model Competition

10 models were built using `caret` with cross-validation (CV), of 5 folds and 3 repetitions, a control parameter and tuning grid, optimized by minimizing LogLoss and evaluating for accuracy. Training was done on normalized, one-hot, resampled data, with feature reduction throughout.

Logistic Regression (all terms, significant terms, interactions, PCA) CV was used on all 4 GLM models. The PCA model was the weakest, while the model with interaction terms and feature reduction model was the strongest, even outperforming Elastic Net.

Elastic Net (all terms, PCA) CV was used to find the optimal lambda & alpha coefficients. The PCA model was the weakest, while the all-terms model was the strongest.

Random Forest (RF) and Gradient Boosting Machine (GBM) For RF, CV was used to find the optimal random feature sampling at each split. For GBM, CV was used to find the optimal number of trees, depth, shrinkage, and nodes. RF outperformed GBM, ElasticNet, and GLM. However, RF took the longest time to train.

Extreme Gradient Boosting (XGB) 5 iterative tuning grids with CV were applied to heuristically comb for 7 optimal hyper-parameters (6480 combinations). Each iteration gradually converged on the following: nrounds 250, depth 5, eta 0.05, gamma 0, subsampling 1, min child 1, colsample 0.6. XGB with feature reduction attained the top accuracy and LogLoss. XGB is the least explainable, requiring ~65 min to train.

Table 2: Model Competition Metrics

Models	CV_Acc	CV_LogLoss
Logistic Regression (all)	0.81758	0.40269
Logistic Regression (sig)	0.83108	0.39754
Logistic Regression (sig+int)	0.86292	3.29989
Logistic Regression (PCA)	0.75101	0.50088
Elastic Net (all)	0.82369	0.39768
Elastic Net (sig)	0.64540	0.42155
Elastic Net (PCA)	0.62574	0.50064
Random Forest (all)	0.93208	0.19474
Random Forest (sig)	0.93066	0.18694
Gradient Boosting (all)	0.90501	0.17136
Gradient Boosting (sig)	0.90485	0.16607
Extreme Gradient (all)	0.94756	0.14089
Extreme Gradient (sig)	0.94818	0.13494

Final Model Selection

The final XGB model with feature reduction was selected with a CV accuracy $> 94\%$ and LogLoss $< .14$. Validation metrics were ~89% accuracy and ~0.27 LogLoss.

Table 3: Predicted Cancellations

Probability	Label
0.7908153	Yes
0.1075948	No
0.0237029	No
0.0495057	No
0.0551797	No
0.9549145	Yes