Telecom Customer Churn Analysis

The BRICS

Introduction - Topic

Prescriptive Analysis

- To uncover the key **cause-and-effect relationships** and understand why
- To **manipulate these factors** in one's favor to get satisfactory outcome

Telecom Customer Churn

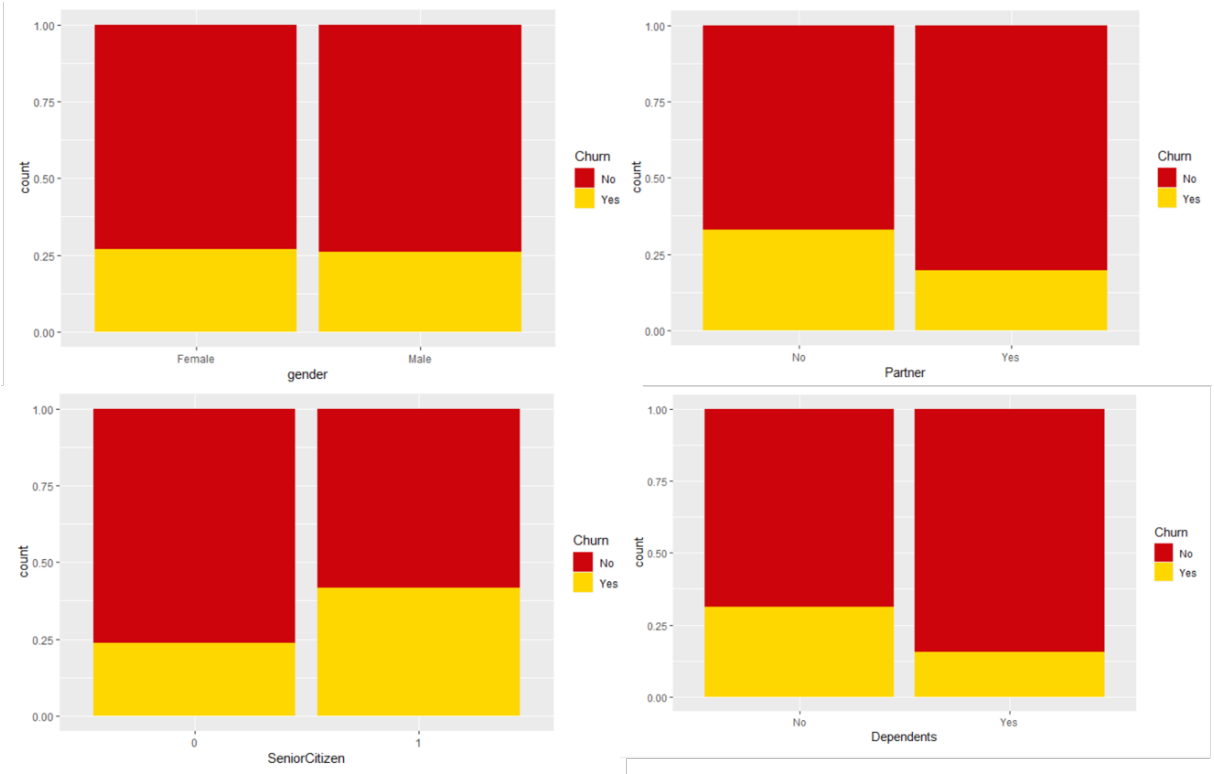
- To find **the factors that result in customer leaving** the company and analyze the reasons
- To give **suggestions on improvement** in order to retain customers

Introduction - Data

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No
2	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes
3	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No
4	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes
5	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No
6	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes

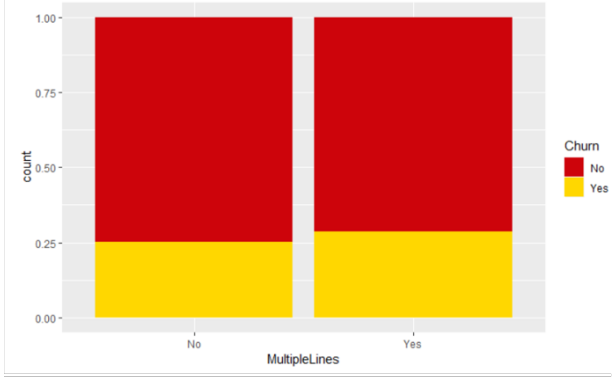
	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
1	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
2	No	No	No	One year	No	Mailed check	56.95	1889.50	No
3	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
4	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
5	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
6	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.50	Yes

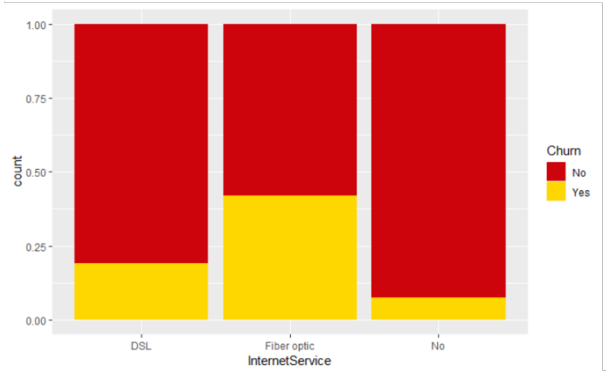
Visualization

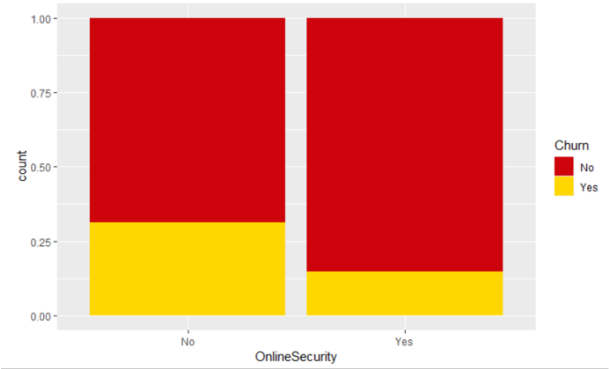


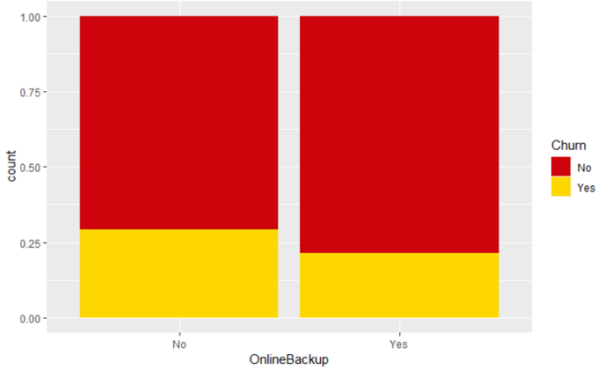
- Gender (X)
- Partner -
- Senior Citizen +
- Dependents -

Visualization (continued)



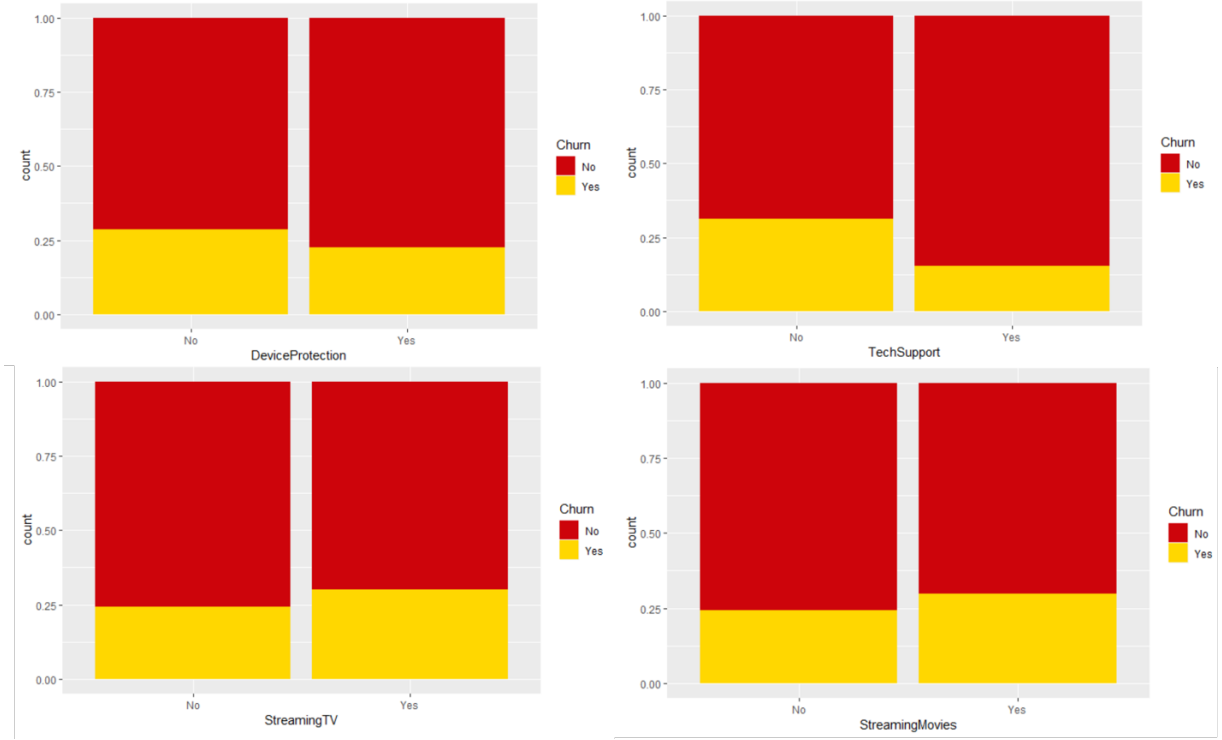






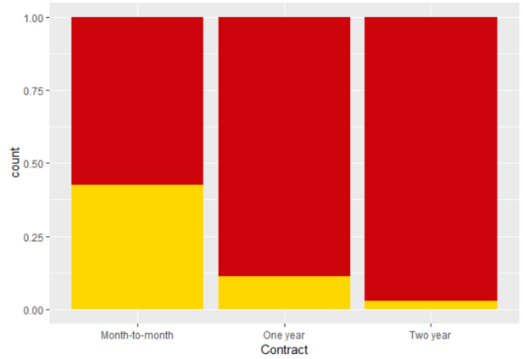
- Multiple Lines +
- Internet Service +
- Online Security -
- Online Backup -

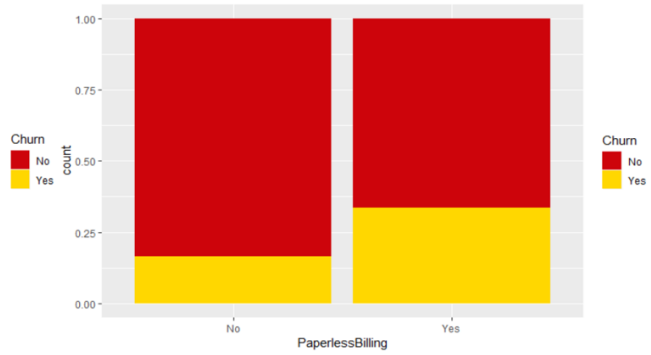
Visualization (continued)

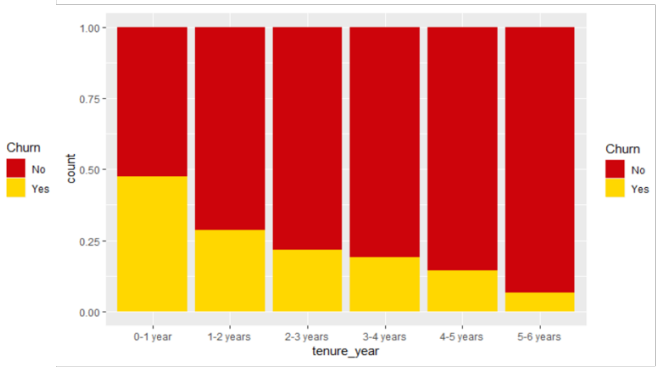


- Device Protection -
- Tech Support -
- Streaming TV +
- Streaming Movies +

Visualization (continued)







- Contract
- Paperless Billing +
- Payment Method
- Tenure -

Data Quality

Generally data is sufficient and clean

Online security: Yes, No, No internet service → Yes, No

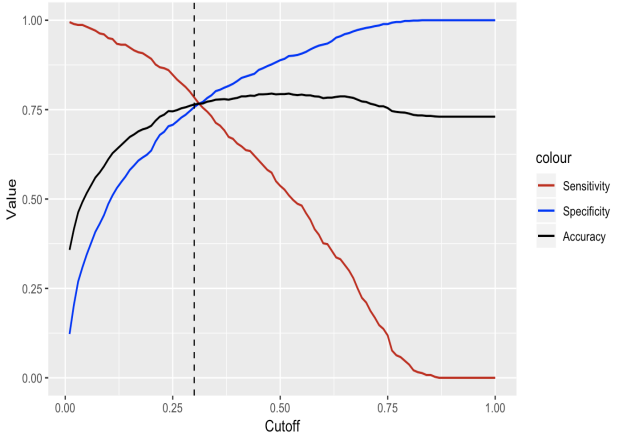
Data Pre-processing

1. Fill NA in total charge (0.15%) with rpart model
2. Replace ‘No internet service’ with ‘No’ in Internet related features
3. Scale numerical features
4. Turn categorical features into dummies

Modeling

1. Models: Logit, knn, tree, random forest, nb, qda, svm
2. Cross-validation: numbers = 5 to avoid overfitting
3. Metrics: emphasis more on Sens than Spec (prob cutoff)

   cost of retain existing customer < attracting new customers



Digression: Best cutoff in business

	Churn	Not churn
Predict positive	**TP**	**FP** (extra cost to retain)
Predict negative	**FN** (extra cost to attract new customer)	**TN**

Quantitative way to get cutoff:

$$Total\ cost = TP \times C_1 + FP \times C_1 + FN \times C_2$$

$$= C_1(TP + FN) \times sen$$

$$+ C_1(TN + FP) \times (1 - spe)$$

$$+ C_2(TP + FN) \times (1 - sen)$$

$$= (C_1 - C_2)(TP + FN) \times sen$$

$$- C_1(TN + FP) \times spe + constant$$

C1 – cost of retain existing customers

C2 – cost of attracting new customers

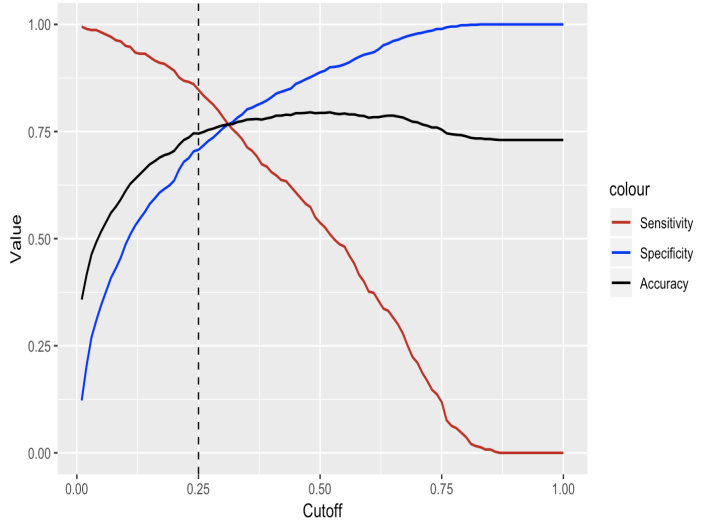
Model Selection

Model Selection

Logistic

Criteria: Searching For Lowest AIC with function stepAIC
Pros: easy to interpret, helpful to determine the quantitative impact of variables
Cutoff Point: At 25%, good sensitivity with not low accuracy and specificity

ROC	Sens	Spec
0.8456	0.8977	0.5507



Model Selection

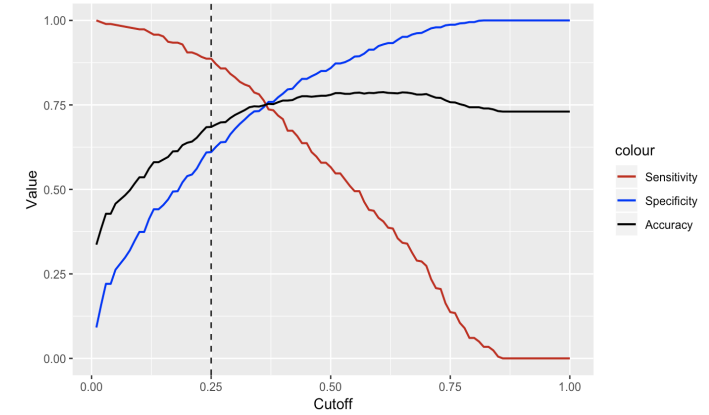
KNN

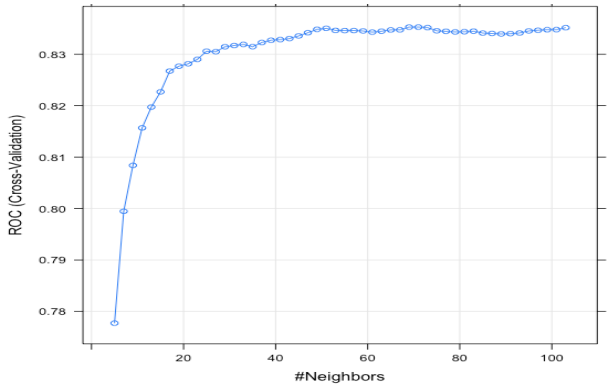
Criteria of Choosing K: Highest AUC when K=71

Pros: Few parameters needed, a good benchmark

Larger cost: a larger drop of accuracy for getting a higher sensitivity than logistic model

Cutoff Point: At 25%, a little higher sensitivity than Logit



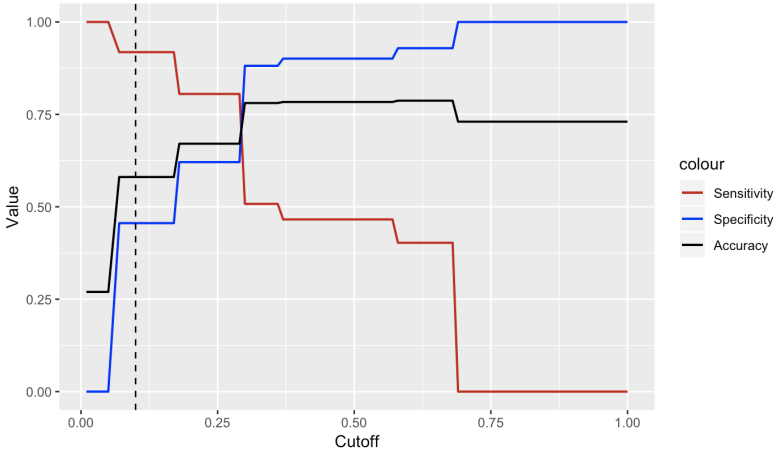


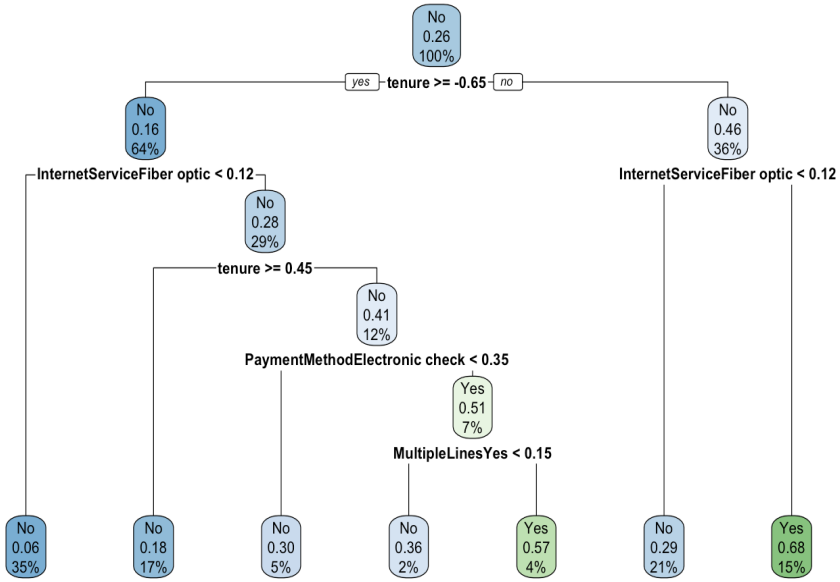
Decision Tree

Criteria: Complexity for 0.0054. Max depth is set to 6. Easy for explaining and avoid overfitting.

Pros: deal with both discrete and continuous variables; easy to understand and interpret

Cutoff Point: At 10%, very costly to get high sensitivity, not better than KNN



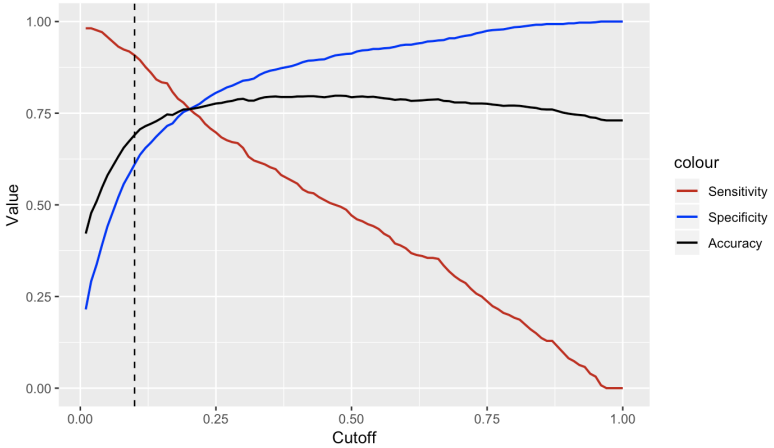


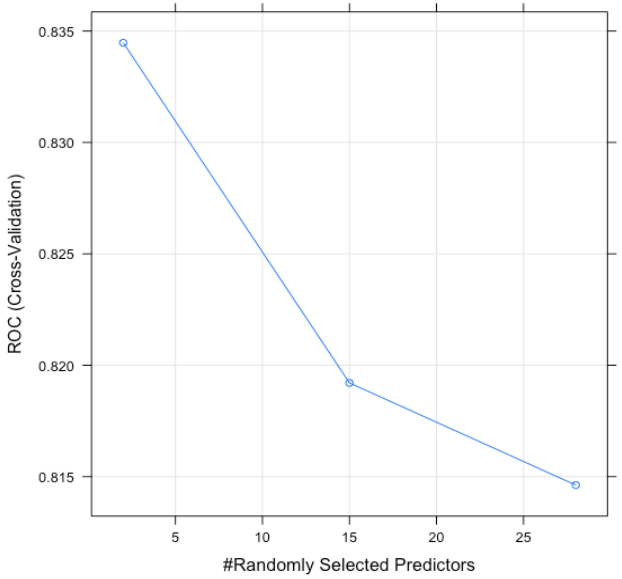
cp	ROC	Sens	Spec
0.0054	0.7785	0.9163	0.4446
0.0062	0.7259	0.9283	0.4137
0.1048	0.6177	0.9573	0.2311

Random Forest

Criteria: Highest accuracy with 500 trees and 2 Maximum Variables at Each Split

Cutoff Point: At 10%, better than the tree, similar to KNN





mtry	ROC	Sens	Spec
2	0.8345	0.9226	0.4601
15	0.8192	0.8938	0.4909
28	0.8146	0.8902	0.4963

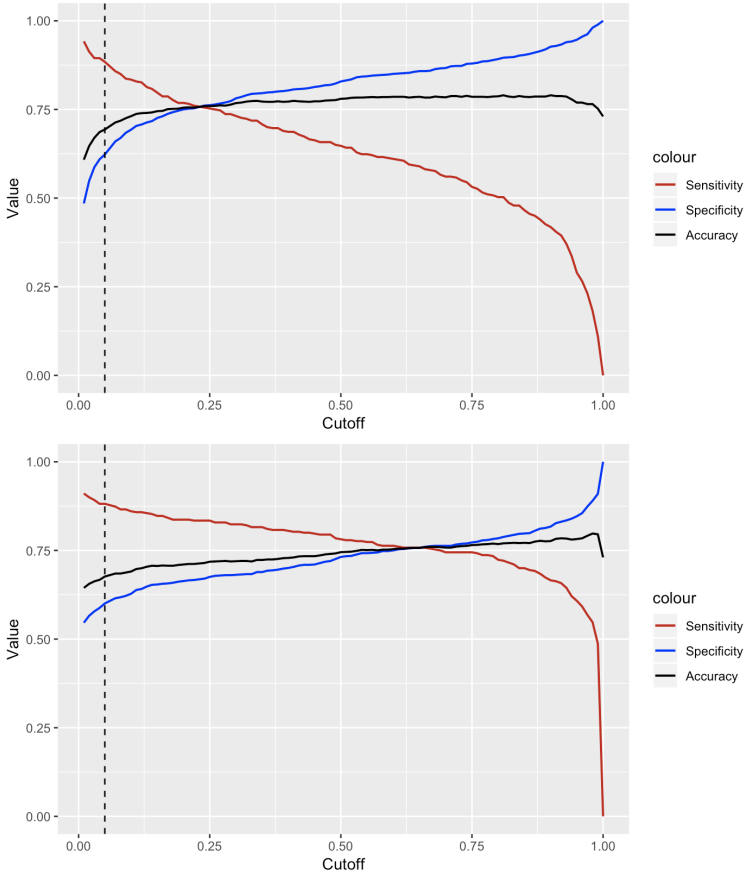
Naive Bayes & Quadratic Discriminant Analysis

Naive Bayes: cost for getting a higher sensitivity is low, but sensitivity not high enough

Cons: needs estimation of many hidden variables, no regularization of overfitting

Quadratic Discriminant Analysis: Similar to NB, even lower cost but also lower available sensitivity

Cutoff Point: At 5%, high sensitivity with not low accuracy and specificity

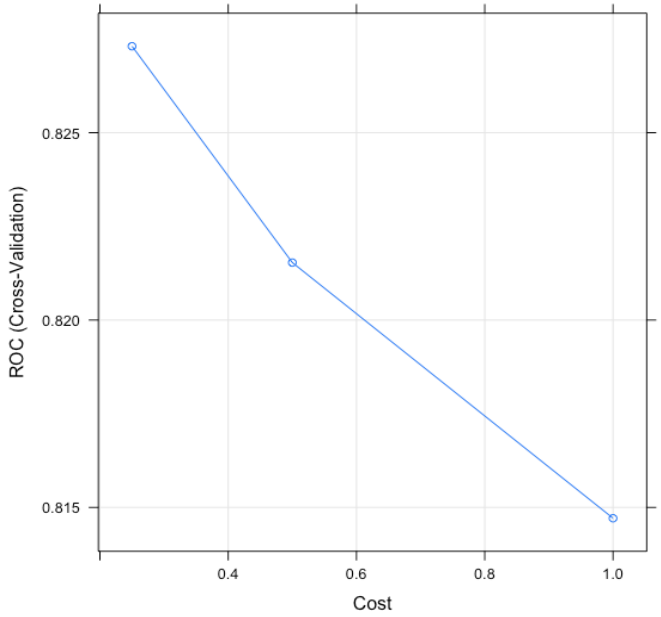


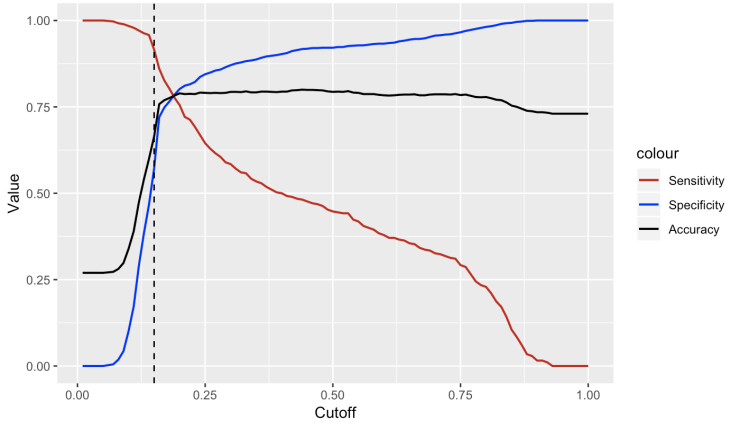
Radial Support Vector Machine

Criteria: Highest AUC (small C = 0.25, which avoids overfitting)

Pros: Low risk for overfitting, perform very well with appropriate kernel function

Cutoff Point: At the elbow of 15%, high sensitivity with not low accuracy and specificity

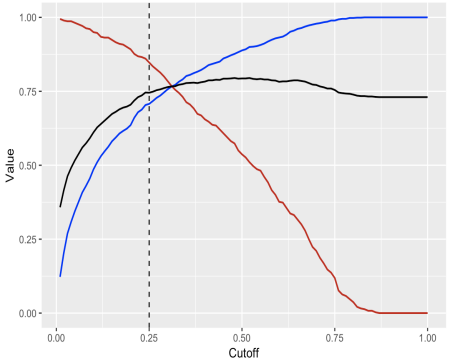




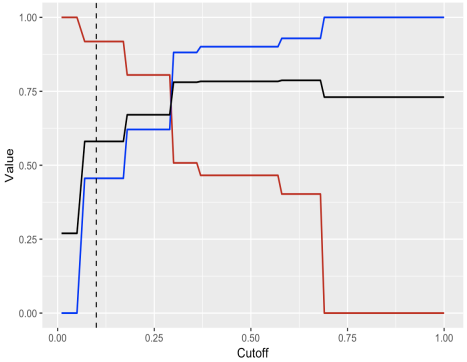
C	ROC	Sens	Spec
0.25	0.8273	0.9273	0.4419
0.5	0.8215	0.9273	0.4419
1	0.8147	0.9250	0.4510

Model Selection

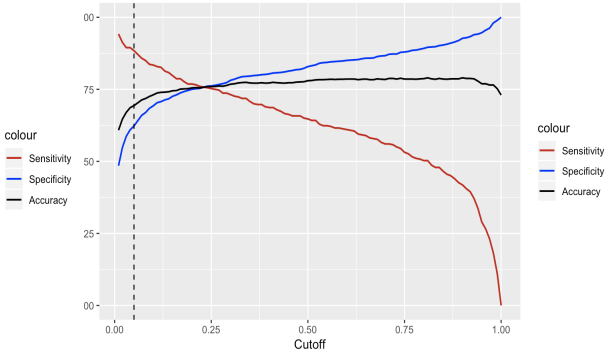
Logit



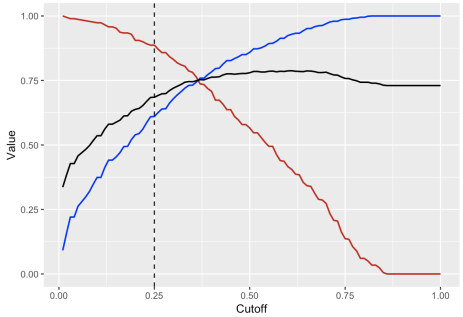
Decision Tree



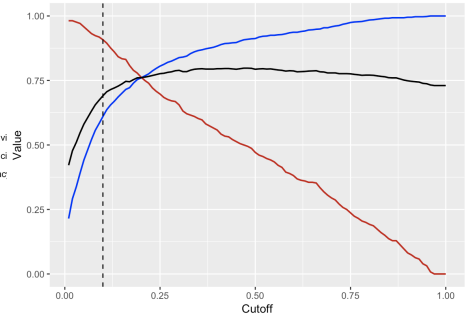
Naive Bayes



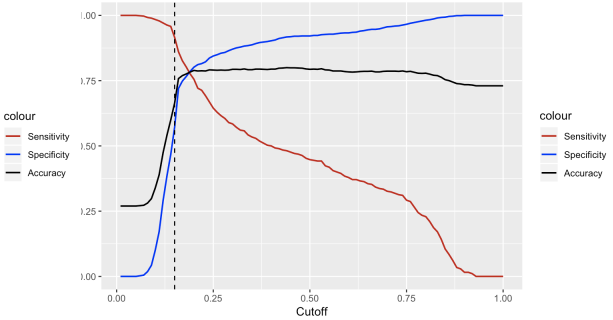
KNN



Random Forest



Support Vector Machine



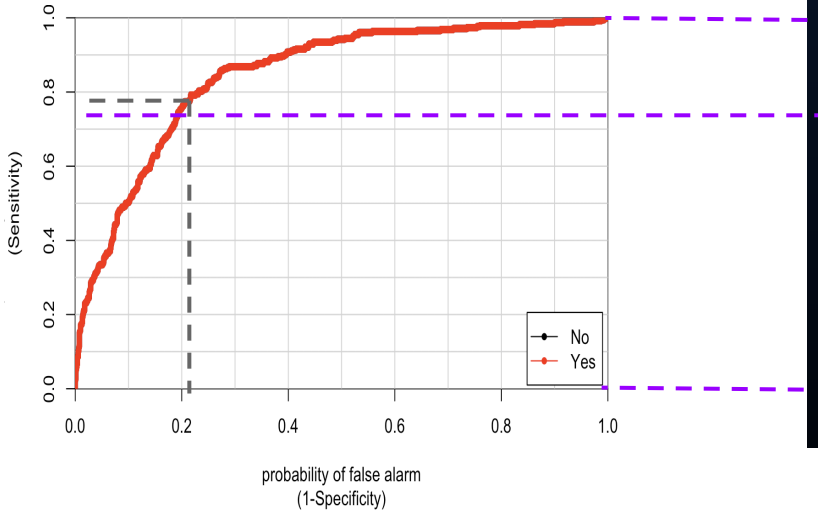
Comparison with DataRobot

DataRobot Best Model (Cross Validated AUC)

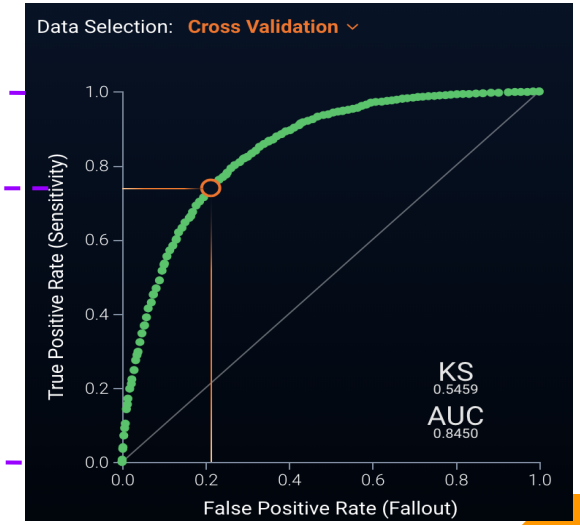
Light Gradient Boosted Trees Classifier with Early Stopping

AUC of best model: 0.8450

ROC Curve of SVM Model



ROC Curve of Light GBM Model

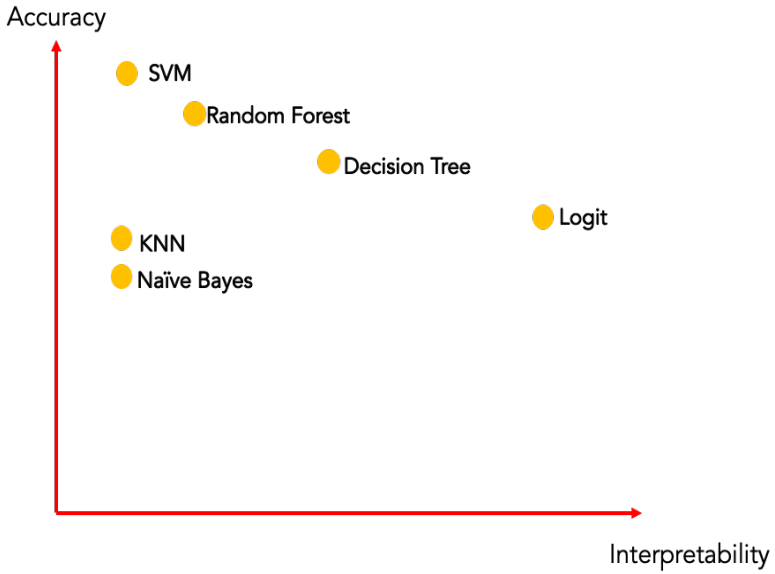


Best Model – Trade-offs

Best Model: Radical Support Vector Machine

Accuracy vs. Interpretability

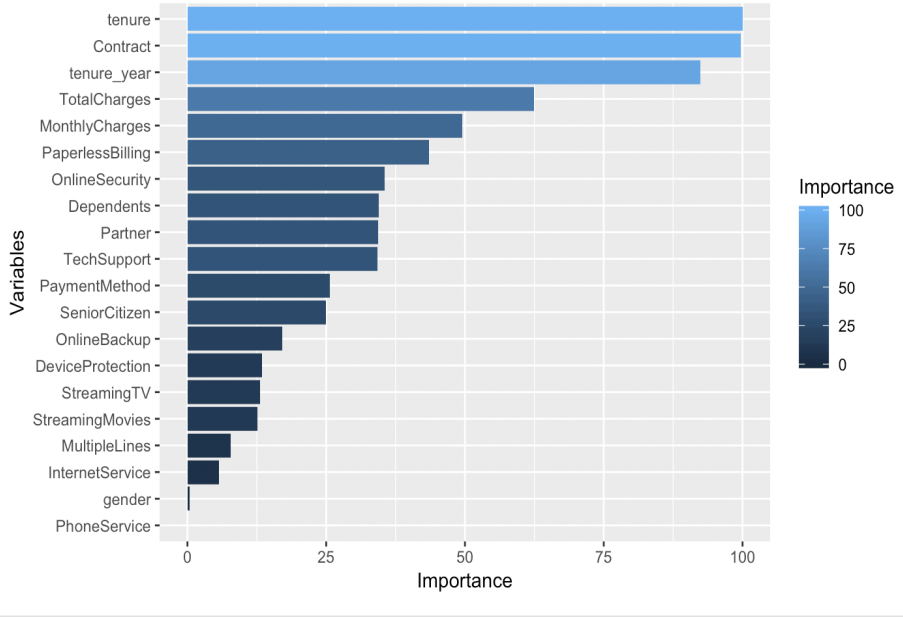
- Higher rate for predicting true positive
- Harder to interpret the detailed relationship between each independent variables and the target variable



Result Interpretation

Rank by Importance Level

In conjunction with Visualization and Logit Coefficients



Result Interpretation

Target variable: Churn

Positive: Total/Monthly Charges; Paperless Billing; Payment Method (Electronic check); Senior Citizen; Streaming TV and Movies; Multiple Lines; Internet Service (Fiber Optics)

Negative: Tenure; Yearly Contract; Online Security; Tech Support; Online Backup; Dependent; Partner; No Internet Service

Business Insight

Problems	Suggestions
Sensitivity of Senior Customer to the monthly charge	Package contract with family
Positive relationship of Paperless Billing	Regular Emails containing: Latest or Special Offers, Usage Report, New Products or Services
More expensive or worse TV, Movie and Internet services	Cooperate with specialized companies Lower Charges
People with Partners More likely to stay	Differentiated Contracts the distinct between single and married
Contract	Lower the price with the increase in the time the contract lasts

Reflection

- Collinearity Between Total Charge & Monthly Charge
- Could Not Quantify Between The Cost of Attracting new Customers & Keeping Current Customers


THANK YOU!