Predicting Customer Churn

Loren Price
Supervised Learning Capstone
Thinkful Data Science Program
7/18/18

Predicting customer churn in the telecommunications Industry.

Losing customers results in direct revenue loss, as well as the expenditure of resources to continually find new customers. Brands across all industries are looking for ways to build meaningful bonds with their customers in order to keep them engaged and loyal to the company. In the telecommunications industry, customer attrition, or churn, is a particularly challenging problem as customers in most markets have access to several options.

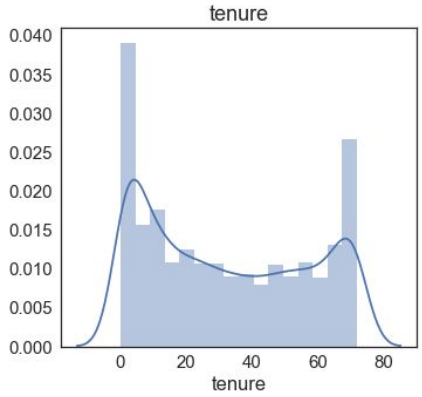
Dataset: Telco Customer Churn

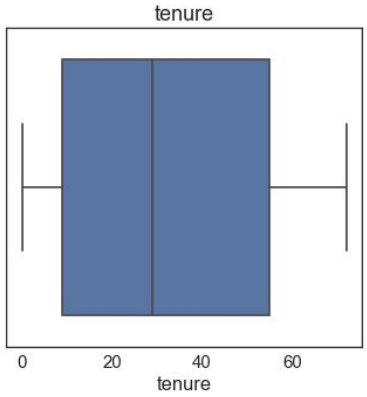
https://www.kaggle.com/blastchar/telco-customer-churn/home

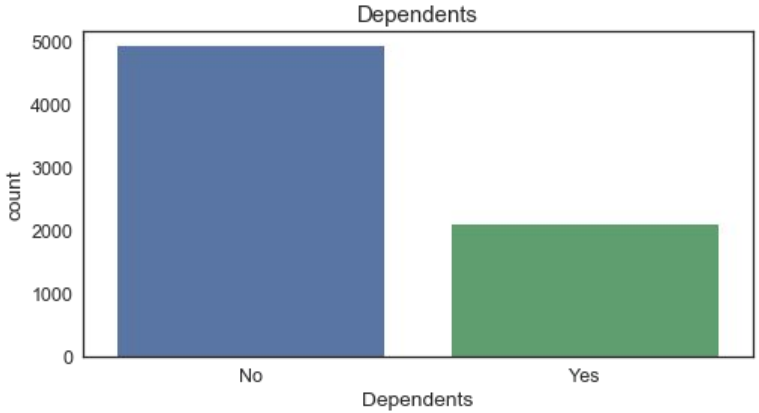
- Published by IBM in 2015.
- The Telco Customer Churn dataset contains customer information from a telecommunications company.
- It includes general demographic information about the customer as well as the various services they were using.
- The raw data contains 7043 rows and 21 columns, with each row being a unique customer.
- The target variable is **Churn**, whether or not the customer left the company in the last 30 days.

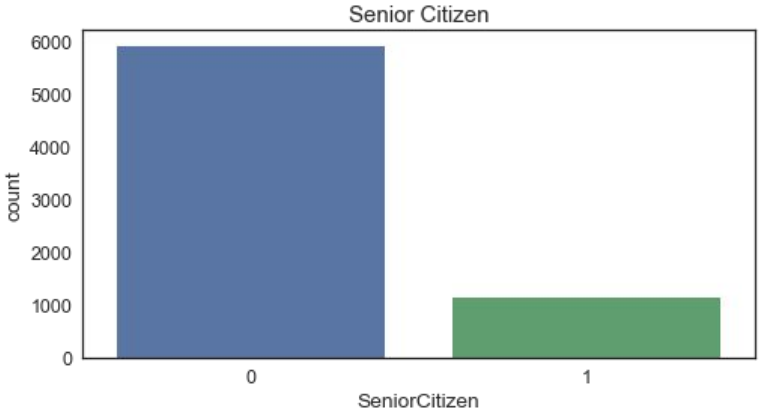
Exploratory Data Analysis

- **Tenure** - How long they have been a customer.
- **Dependents** - If the customer has any dependents.
- **Senior Citizen** - If the customer is a senior citizen.



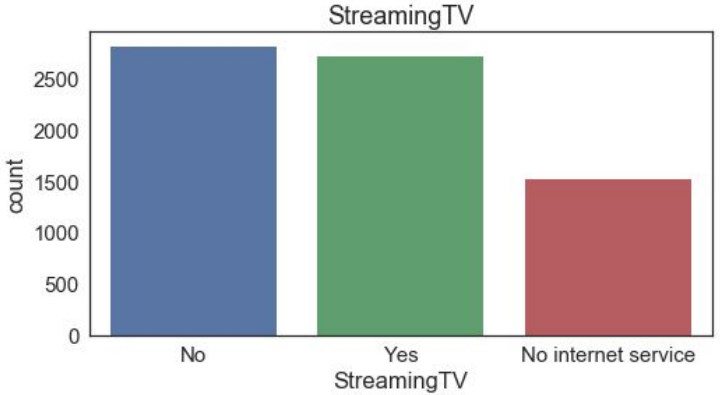


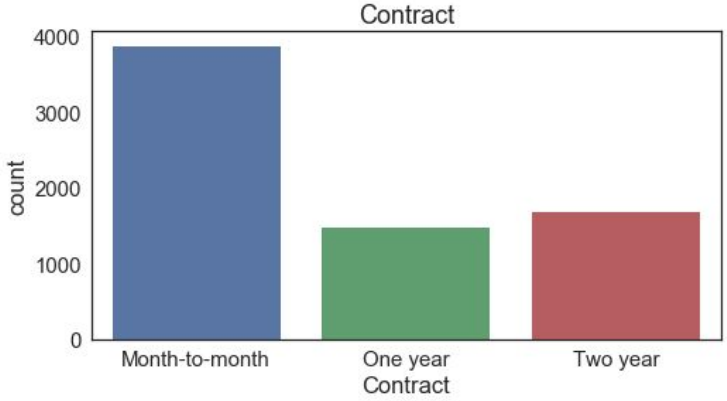


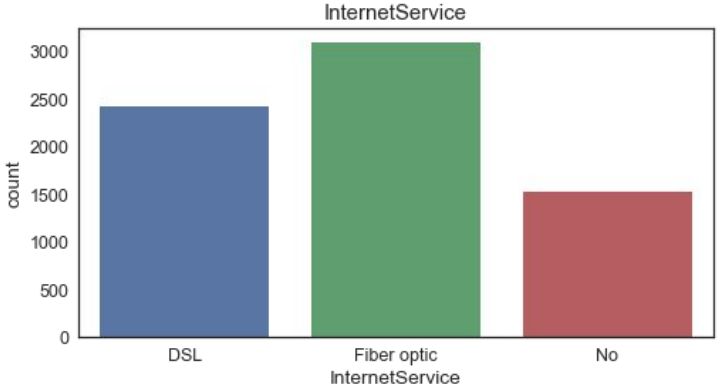


Exploratory Data Analysis

- **StreamingTV** If the customer subscribes to a streaming TV service.
- **Contract** - What type of contract the customer has.
- **InternetService** - What type of internet service the customer has.



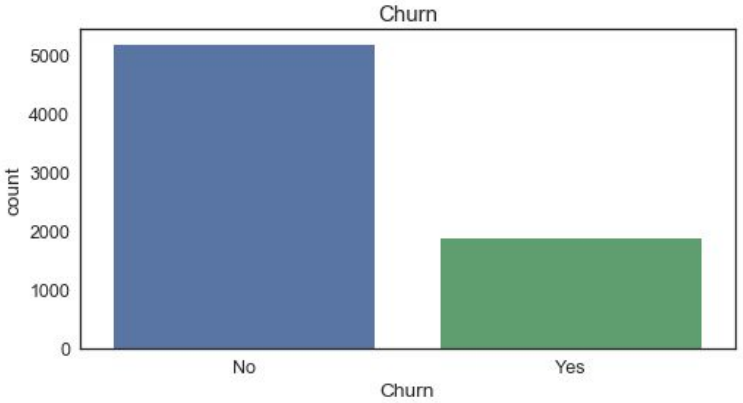


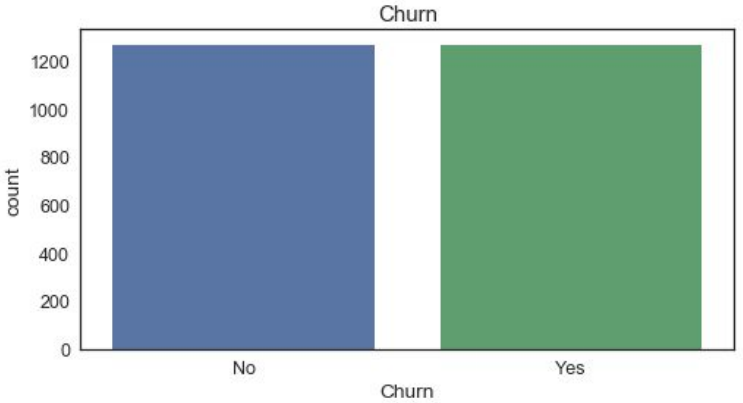


Target Variable - Churn

- **Baseline Accuracy of .76**
  Models were built using full dataset, and best results were .81

- **To combat class imbalance, the dominant class was down-sampled to achieve a .50 baseline.**





Feature Engineering and Selection

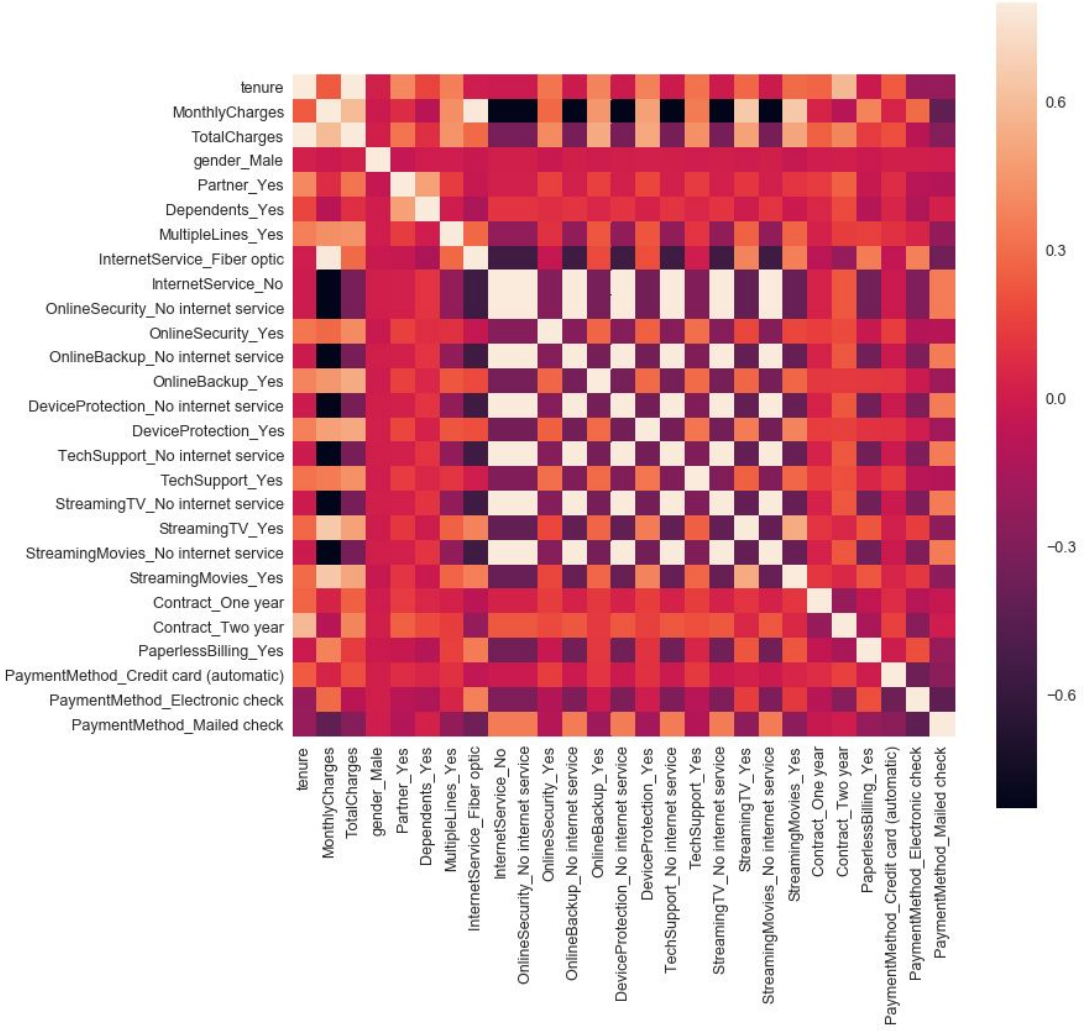
Dummy variables were created for the following features:

- gender
- Partner
- Dependents
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- Contract
- PaperlessBilling
- PaymentMethod

The following features were dropped:

- customerID
  Not related to outcome
- SeniorCitizen
- PhoneService
  Few observations outside of dominant class.

Feature Correlation



Modeling

- Naive Bayes
- K Nearest Neighbors
- Decision Tree
- Random Forest
- Logistic Regression
- Logistic Regression
  - With Ridge (L2) Regularization
  - With Lasso (L1) Regularization
- Support Vector Machine
- Gradient Boosting Classifier

Naive Bayes

- Computationally light.
- Assumes variables are independent from each other.
- Returns a category based on probability.

$$\hat{y} = \underset{y}{argmax}\, P(y) \prod_{i=1}^{n} P(x_i|y)$$

Naive Bayes Results

Cross Validation Mean: **0.7170**
Cross Validation Range: **0.0945**

Sensitivity: **0.8672**
Specificity: **0.5623**

Type 1 error: **165/377**
Type 2 error: **51/384**

K Nearest Neighbors

- Computationally light.
- Based on the distance to other data points. Commonly the euclidean distance.
- K number of points get to vote on which class it belongs to.

$$Dist(X^n, X^m) = \sqrt{\sum_{i=1}^{D} (X_i^n - X_i^m)^2}$$

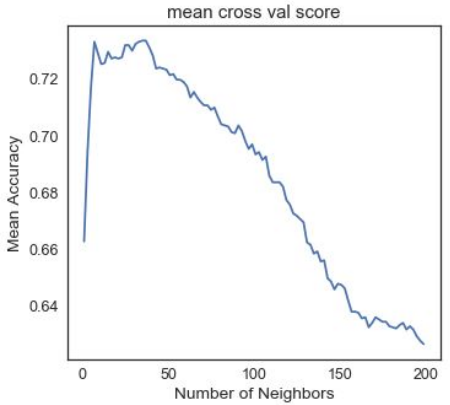
K Nearest Neighbors Results

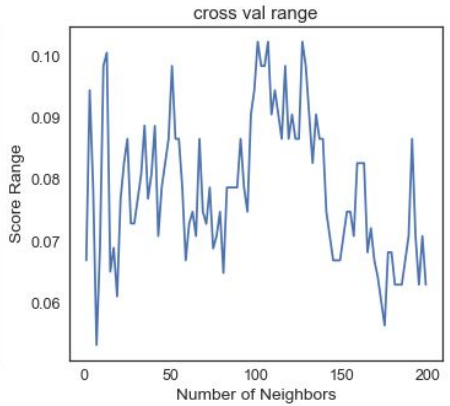
Parameter Settings
n_neighbors: 53
Weights: 'distance'

Cross Validation Mean: **0.7277**
Cross Validation Range: **0.0590**

Sensitivity: **0.7031**
Specificity: **0.7294**

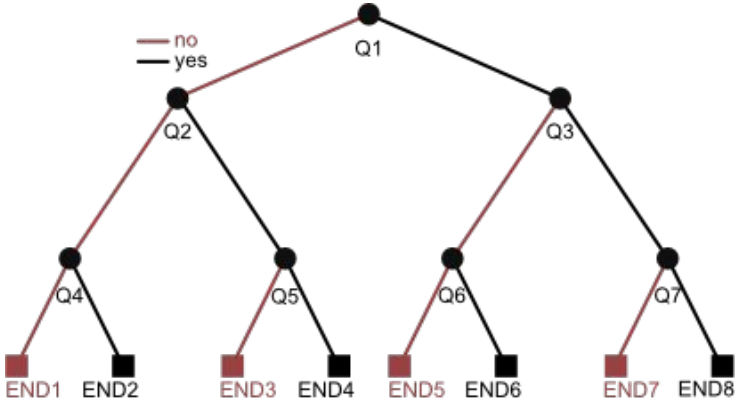
Type 1 error: **102/377**
Type 2 error: **114/384**





Decision Tree

- Computationally light.
- Asks questions to minimize entropy at each step.
- Comprised of nodes that are either questions or leaf nodes (endpoints).
- Requires balanced data.
- Prone to overfitting.



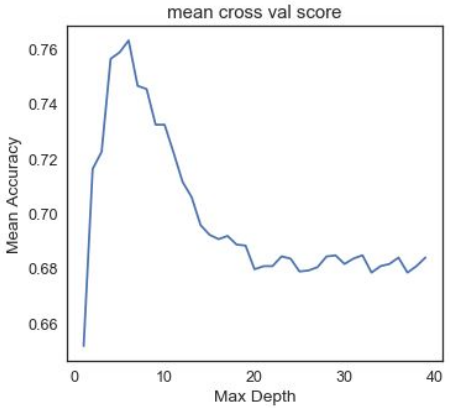
Decision Tree Results

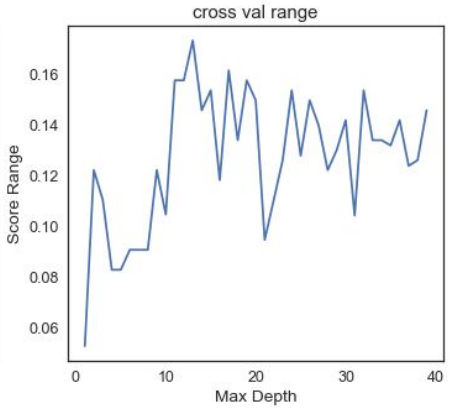
Parameter Settings
max_depth: 3
max_features: 27

Cross Validation Mean: **0.7226**
Cross Validation Range: **0.1102**

Sensitivity: **0.6042**
Specificity: **0.8090**

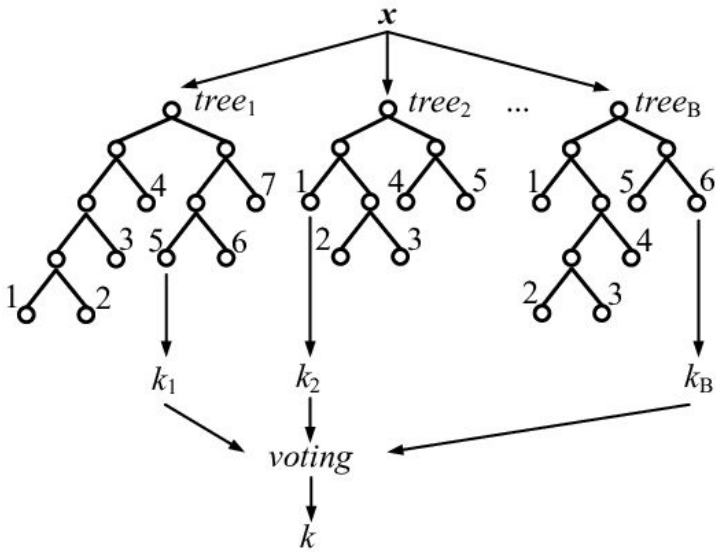
Type 1 error: **72/377**
Type 2 error: **152/384**





Random Forest

- ‘Black box’ model.
- Can be computationally heavy.
- Comprised of multiple decision trees.
- Uses bagging, where each tree selects a subset of observations, with replacement, and runs models in parallel.
- In classification, the mode is returned.



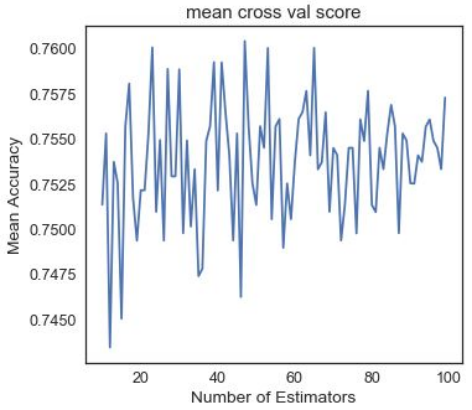
Random Forest Results

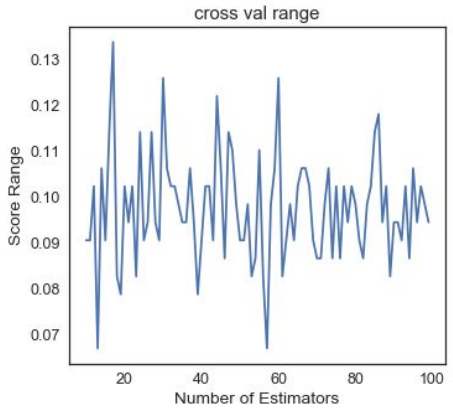
Parameter Settings
n_estimators: 38
max_depth: 7
max_features: 'log2'
min_samples_leaf: 1

Cross Validation Mean: **0.7652**
Cross Validation Range: **0.0906**

Sensitivity: **0.7760**
Specificity: **0.7321**

Type 1 error: **101/377**
Type 2 error: **86/384**





Logistic Regression

- Calculates the probability of a binary outcome variable.
- Uses the natural log of the odds that the target variable equals one of the categories.

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Logistic Regression Results

Cross Validation Mean: **0.7580**
Cross Validation Range: **0.0967**

Sensitivity: **0.7630**
Specificity: **0.7347**

Type 1 error: **100/377**
Type 2 error: **91/384**

	features	coefficients	absolute_coef
0	Contract_Two year	-1.314293	1.314293
1	Contract_One year	-0.678860	0.678860
2	TechSupport_Yes	-0.506970	0.506970
3	PaymentMethod_Credit card (automatic)	-0.418411	0.418411
4	PaperlessBilling_Yes	0.361630	0.361630
5	Partner_Yes	-0.344876	0.344876
6	InternetService_Fiber optic	0.305331	0.305331
7	OnlineBackup_Yes	-0.199402	0.199402
8	PaymentMethod_Electronic check	0.175669	0.175669
9	OnlineSecurity_Yes	-0.166100	0.166100
10	StreamingMovies_Yes	0.153785	0.153785
11	PaymentMethod_Mailed check	-0.148340	0.148340
12	DeviceProtection_Yes	-0.102946	0.102946
13	tenure	-0.054814	0.054814
14	TechSupport_No internet service	-0.046116	0.046116
15	OnlineBackup_No internet service	-0.046116	0.046116
16	StreamingTV_No internet service	-0.046116	0.046116
17	StreamingMovies_No internet service	-0.046116	0.046116
18	OnlineSecurity_No internet service	-0.046116	0.046116
19	InternetService_No	-0.046116	0.046116
20	DeviceProtection_No internet service	-0.046116	0.046116
21	StreamingTV_Yes	0.043070	0.043070
22	MultipleLines_Yes	0.043003	0.043003
23	Dependents_Yes	0.021063	0.021063
24	MonthlyCharges	0.018574	0.018574
25	gender_Male	0.007903	0.007903
26	TotalCharges	0.000238	0.000238

Logistic Regression With Ridge (L2) Regularization

- Regularization helps prevent overfitting.
- Adds a penalty to the cost function for large coefficients.
- Penalizes by the sum of squared coefficients.

$$\lambda \sum_{j=0}^{M} W_j^2$$

Ridge Logistic Regression Results

Parameter Settings
C: 0.01

Cross Validation Mean: **0.7502**
Cross Validation Range: **0.0867**

Sensitivity: **0.7266**
Specificity: **0.7480**

Type 1 error: **95/377**
Type 2 error: **105/384**

Logistic Regression With Lasso (L1) Regularization

- Helps prevent overfitting.
- Penalizes by the sum of the absolute value of the coefficients.
- Works as embedded feature selection by reducing small parameter estimates to be equal to 0.

$$\lambda \sum_{j=0}^{M} |W_j|$$

Lasso Logistic Regression Results

Parameter Settings
C: 0.01

Cross Validation Mean: **0.7372**
Cross Validation Range: **0.0574**

Sensitivity: **0.7786**
Specificity: **0.6631**

Type 1 error: **127/377**
Type 2 error: **85/384**

	features	coefficients	absolute_coef
0	tenure	-0.090061	0.090061
1	MonthlyCharges	0.021067	0.021067
2	TotalCharges	0.000457	0.000457
3	TechSupport_No internet service	0.000000	0.000000
4	PaymentMethod_Electronic check	0.000000	0.000000
5	PaymentMethod_Credit card (automatic)	0.000000	0.000000
6	PaperlessBilling_Yes	0.000000	0.000000
7	Contract_Two year	0.000000	0.000000
8	Contract_One year	0.000000	0.000000
9	StreamingMovies_Yes	0.000000	0.000000
10	StreamingMovies_No internet service	0.000000	0.000000
11	StreamingTV_Yes	0.000000	0.000000
12	StreamingTV_No internet service	0.000000	0.000000
13	TechSupport_Yes	0.000000	0.000000
14	DeviceProtection_No internet service	0.000000	0.000000
15	DeviceProtection_Yes	0.000000	0.000000
16	OnlineBackup_Yes	0.000000	0.000000
17	OnlineBackup_No internet service	0.000000	0.000000
18	OnlineSecurity_Yes	0.000000	0.000000
19	OnlineSecurity_No internet service	0.000000	0.000000
20	InternetService_No	0.000000	0.000000
21	InternetService_Fiber optic	0.000000	0.000000

Feature Selection Using Lasso Coefficients

Features Used:

- **tenure**
- **MonthlyCharges**
- **TotalCharges**

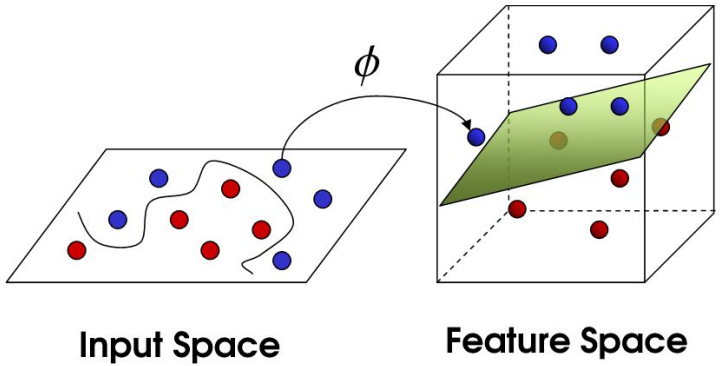
Cross Validation Mean: **0.7372**
Cross Validation Range: **0.0574**

- **Accuracy remained at .7372 with .0574 cross validation range after dropping all but 3 variables.**

	features	coefficients	absolute_coef
0	tenure	-0.090061	0.090061
1	MonthlyCharges	0.021067	0.021067
2	TotalCharges	0.000457	0.000457
3	TechSupport_No internet service	0.000000	0.000000
4	PaymentMethod_Electronic check	0.000000	0.000000
5	PaymentMethod_Credit card (automatic)	0.000000	0.000000
6	PaperlessBilling_Yes	0.000000	0.000000

Support Vector Machine

- Creates a boundary between classes.
- Optimizes the margin between the nearest data points and the boundary.
- The cost function balances the size of the margin with the points on the wrong side of the margin.
- Uses the ‘kernel trick’ to find the boundary.



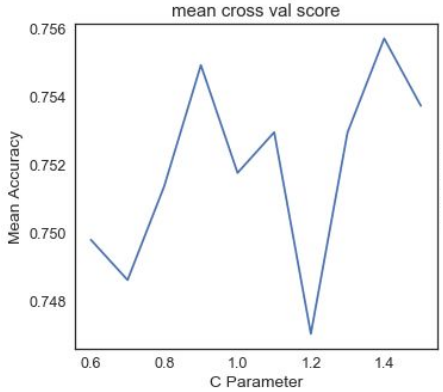
Support Vector Machine Results

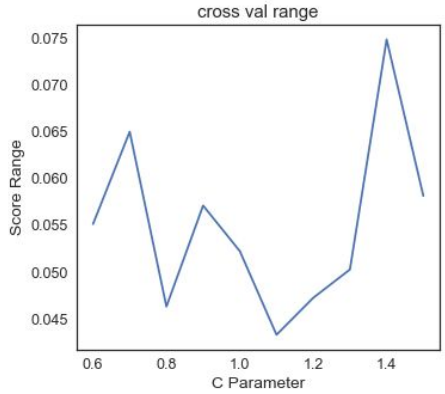
Parameter Settings
kernel: 'linear'
C: 1.1

Cross Validation Mean: **0.7530**
Cross Validation Range: **0.0433**

Sensitivity: **0.6615**
Specificity: **0.8037**

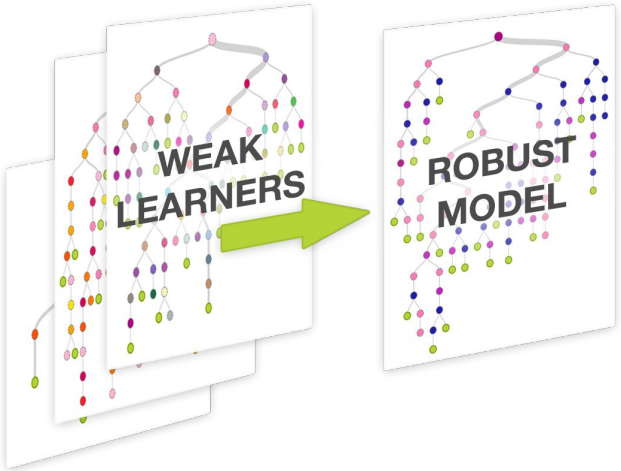
Type 1 error: **74/377**
Type 2 error: **130/384**





Gradient Boosting Classifier

- Models the data sequentially, and adjusts each time based on what was learned.
- Starts with one ‘weak learning’ tree, and subsequent trees are modeled on the data that was not correctly predicted.



Gradient Boosting Results

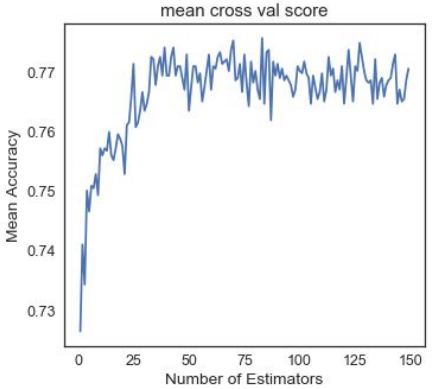
Parameter Settings

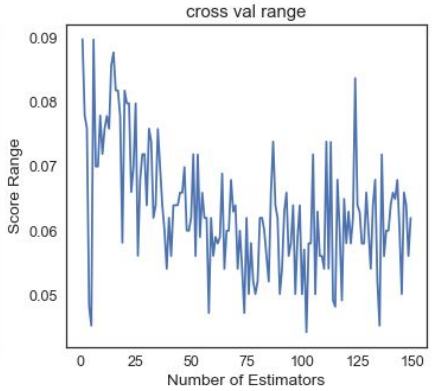
learning_rate: 0.1
max_depth: 8
min_samples_leaf: 21
subsample: 0.75
n_estimators: 67
min_samples_split: 500
max_features: 'sqrt'

Cross Validation Mean: **0.7735**
Cross Validation Range: **0.0945**

Sensitivity: **0.7656**
Specificity: **0.7719**

Type 1 error: **86/377**
Type 2 error: **90/384**





Principal Component Analysis + Gradient Boosting

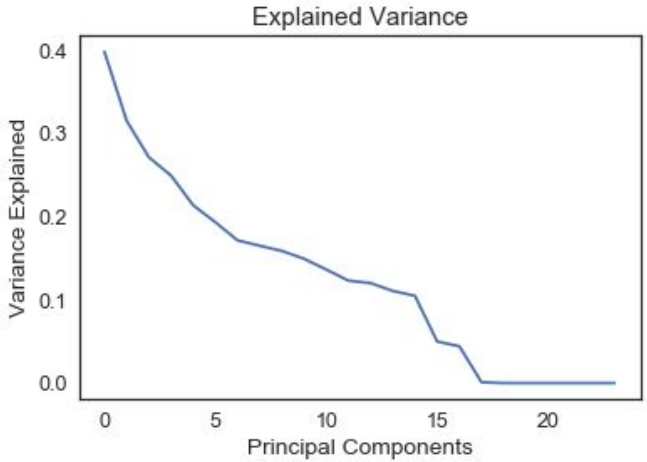
Parameter Settings - *Number of components: 2*

learning_rate: 0.1	n_estimators: 67
max_depth: 8	min_samples_split: 450
min_samples_leaf: 42	max_features: 'sqrt'
subsample: 0.7	

Cross Validation Mean: **0.7486**
Cross Validation Range: **0.0591**

Sensitivity: **0.7396**
Specificity: **0.7586**

Type 1 error: **91/377**
Type 2 error: **100/384**



Model Results

Model	Cross Val Mean	Cross Val Range	Type 1 Error	Type 2 Error	Sensitivity	Specificity	Run Time - Seconds
naive bayes	0.7170384952	0.09448818898	165	51	0.8671875	0.5623342175	0.01
knn	0.7277215348	0.05905511811	102	114	0.703125	0.7294429708	0.02
decision tree	0.7225753031	0.1102362205	72	152	0.6041666667	0.8090185676	0.01
random forest	0.7652074741	0.0905511811	101	86	0.7760416667	0.7320954907	0.08
logistic regression	0.7580458693	0.09670666167	100	91	0.7630208333	0.7347480106	0.02
ridge logistic reg	0.7501906012	0.08661417323	95	105	0.7265625	0.7480106101	0.01
lasso logistic reg	0.7371828521	0.05739907512	127	85	0.7786458333	0.6631299735	0.01
svm	0.7529519778	0.04330708661	74	130	0.6614583333	0.8037135279	407.72
gradient boosting	0.7734814398	0.09448818898	86	90	0.765625	0.7718832891	0.1
pca gradient boost	0.748615798	0.05905511811	91	100	0.7395833333	0.7586206897	0.08

Model Results

- Overall, overfitting was an issue, so the models with the lowest cross validation scores were preferred.
- The support vector machine has the best balance of accuracy and low cross validation range. However due to it's extremely long runtime (407 seconds), the best model is the gradient boosting with PCA.