Predicting Telco Customer Churn

Using Supervised Learning Models

Solomon Mengistu

Outline

- Introduction
- Data Analysis and Visualization
- Preparing for Data for Modeling
- Models
- Model Evaluation methods
- Dimensionality Reduction with PCA
- Feature Selection with SelectKBest
- Comparing Models
- Conclusions and Recommendations

The Jupyter notebook used for this project is available in the link below

https://github.com/sollsam/Supervised_Learning_capstone

Introduction

- Customer churn
  - What is churn?
  - Why study churn?
- The goal
  - Who is leaving?
  - Why?

Churn rate, when applied to a customer base, refers to
a given time period. It is a possible indicator of custome

About the data

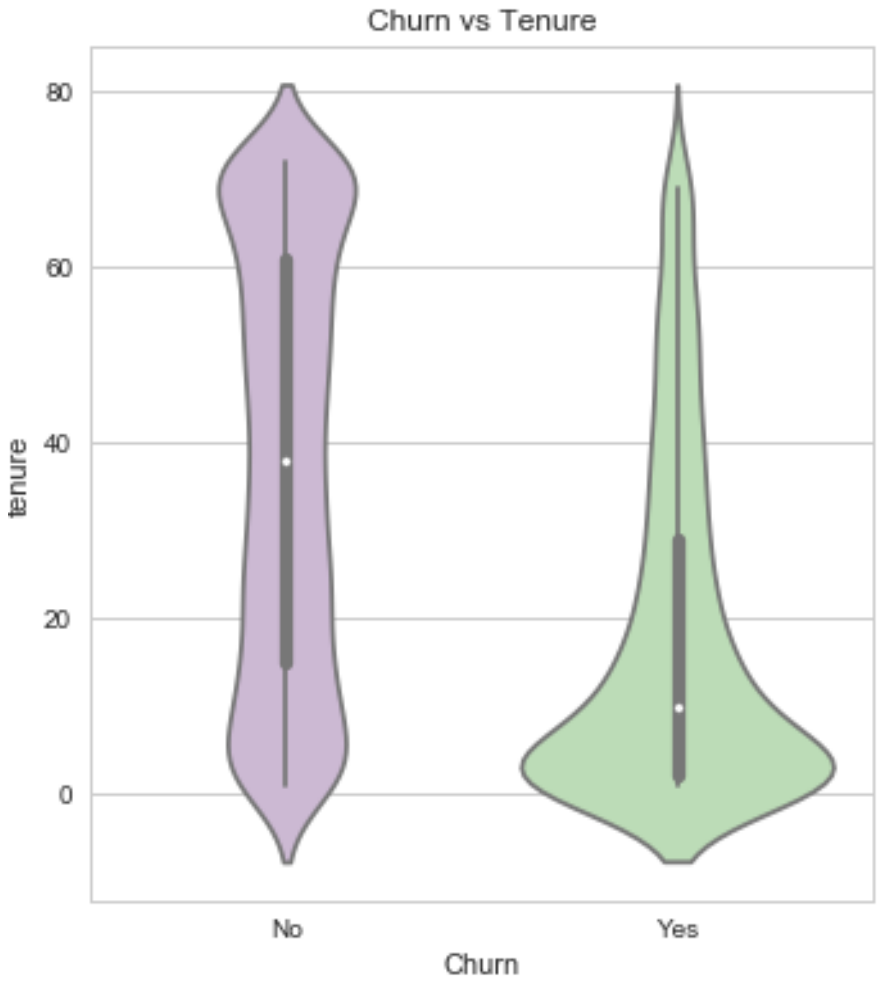
- Shape
- Variables
- Class
- 26.5% Churn

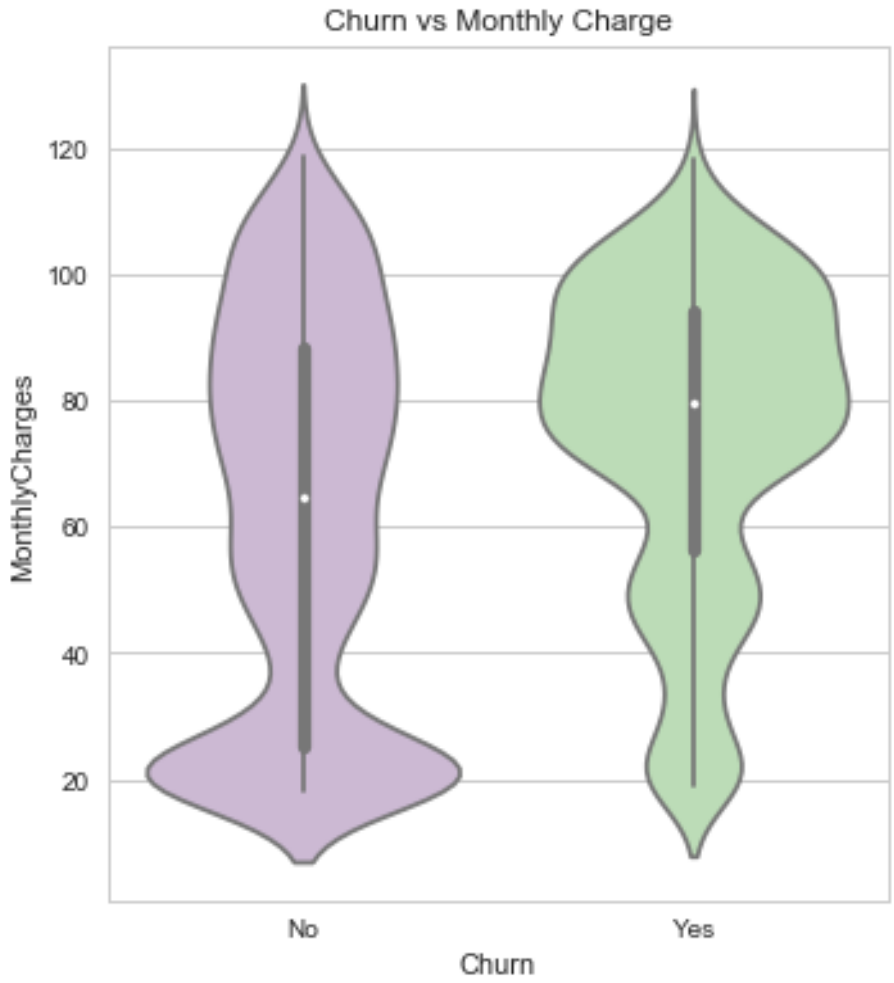
```
No     5163
Yes    1869
Name: Churn, dtype: int64
```

```
data.info()
```

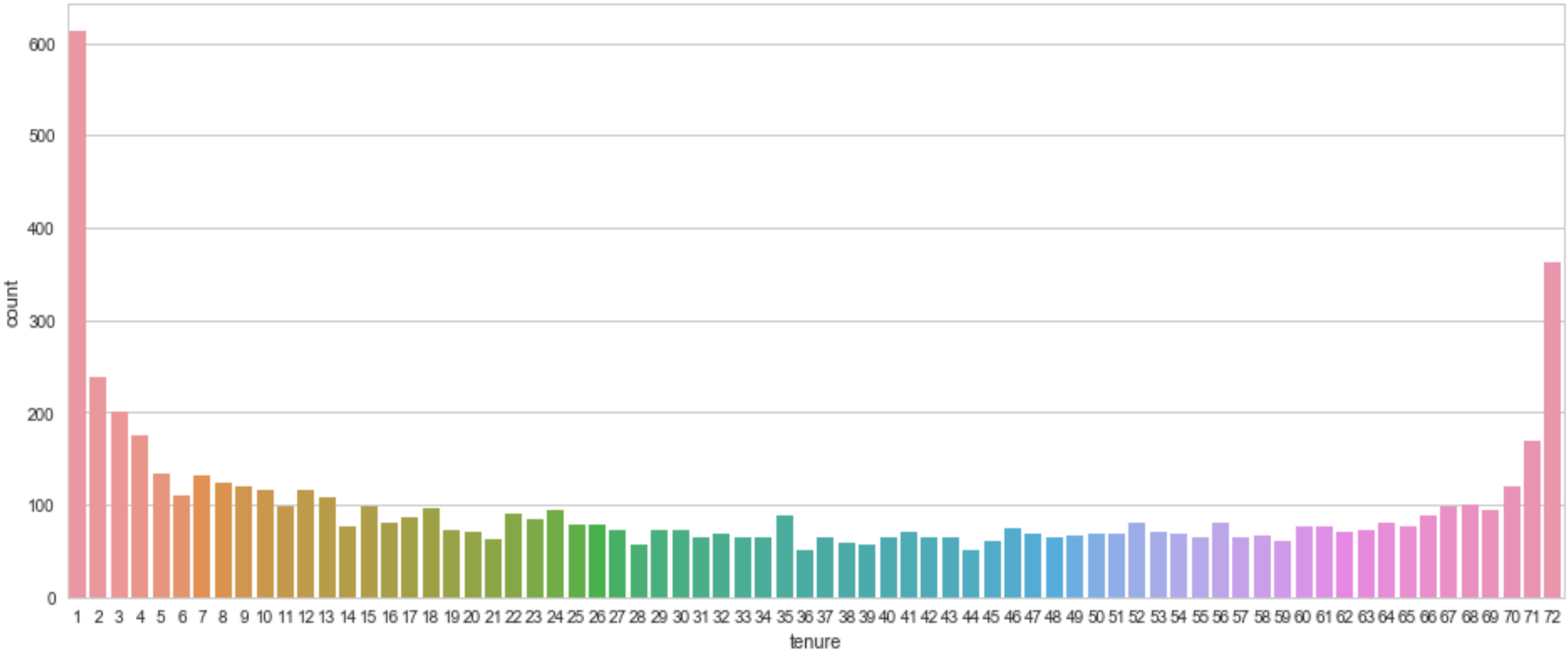
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
customerID          7043 non-null object
gender              7043 non-null object
SeniorCitizen       7043 non-null int64
Partner             7043 non-null object
```

Data Analysis with Visuals

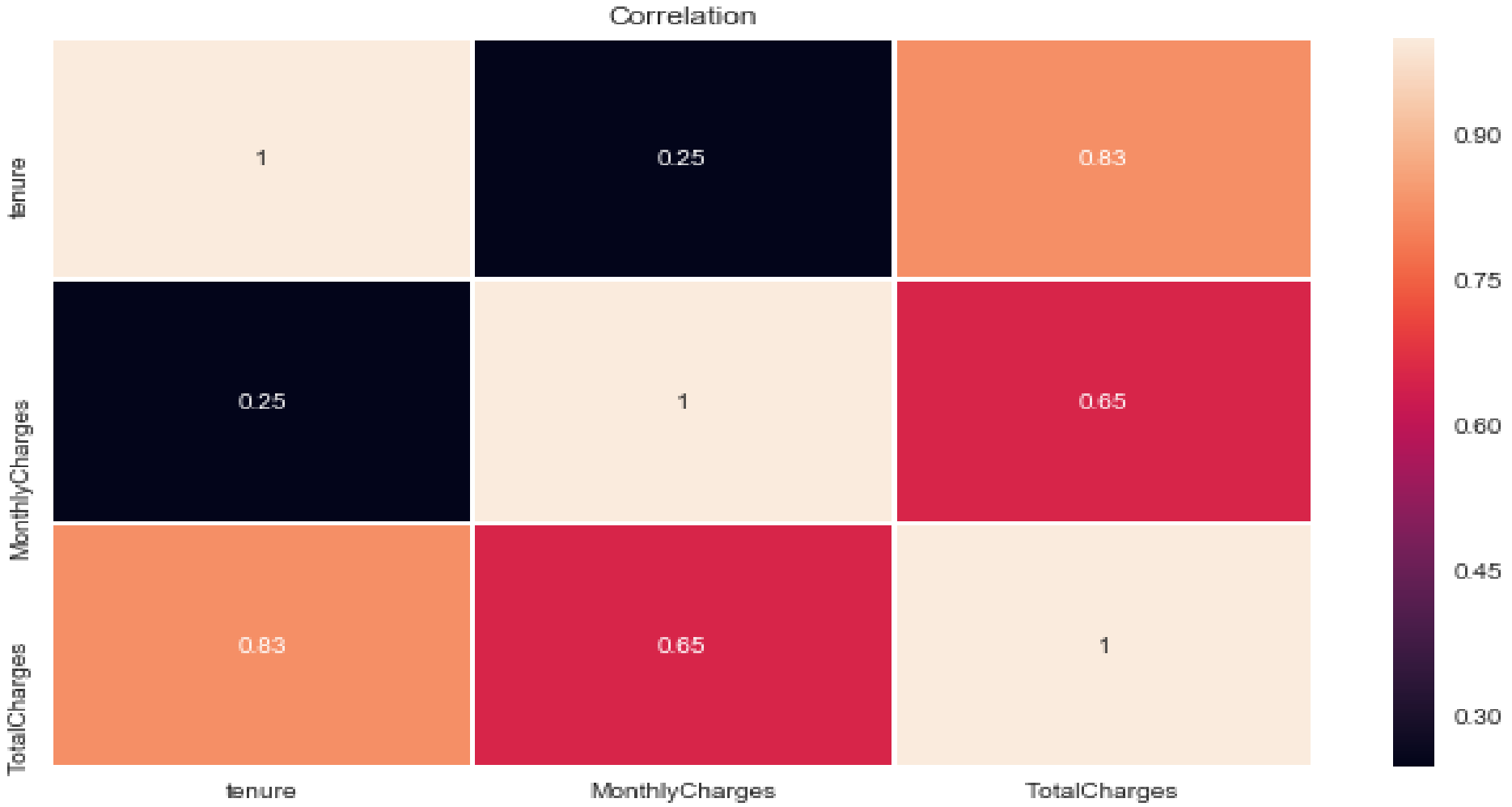




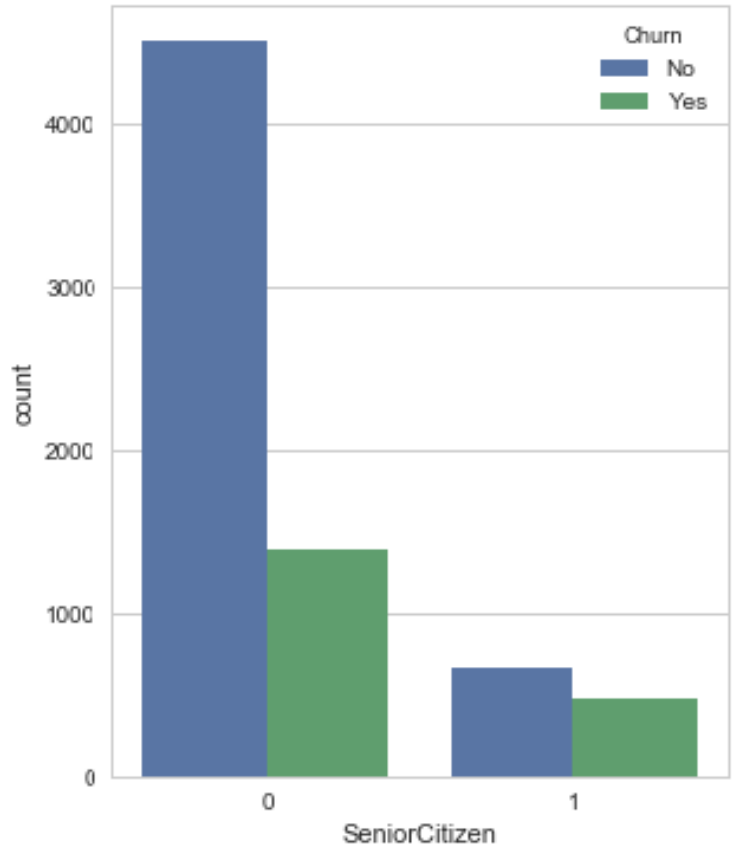
Distribution of tenure

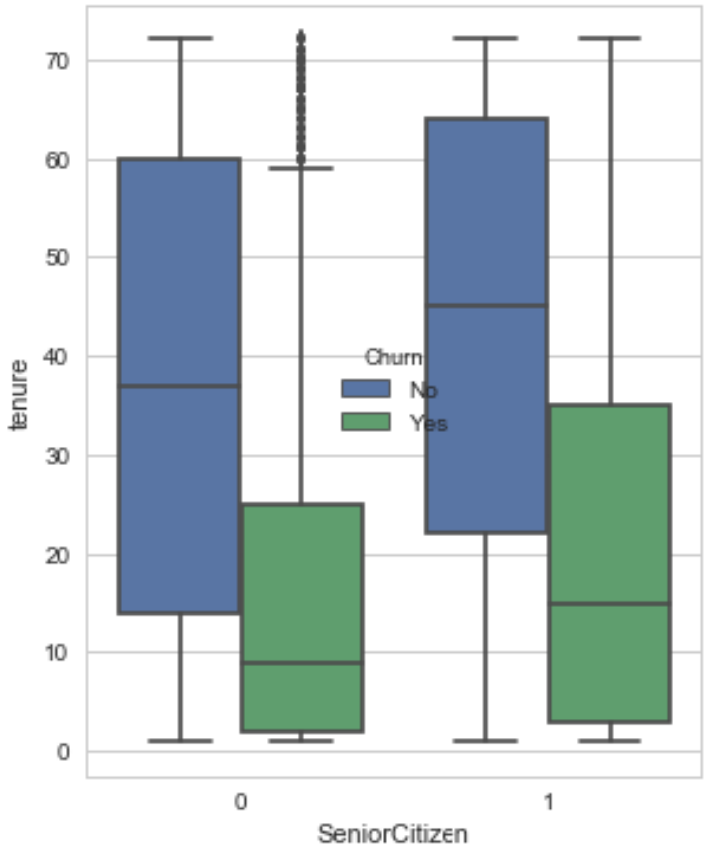


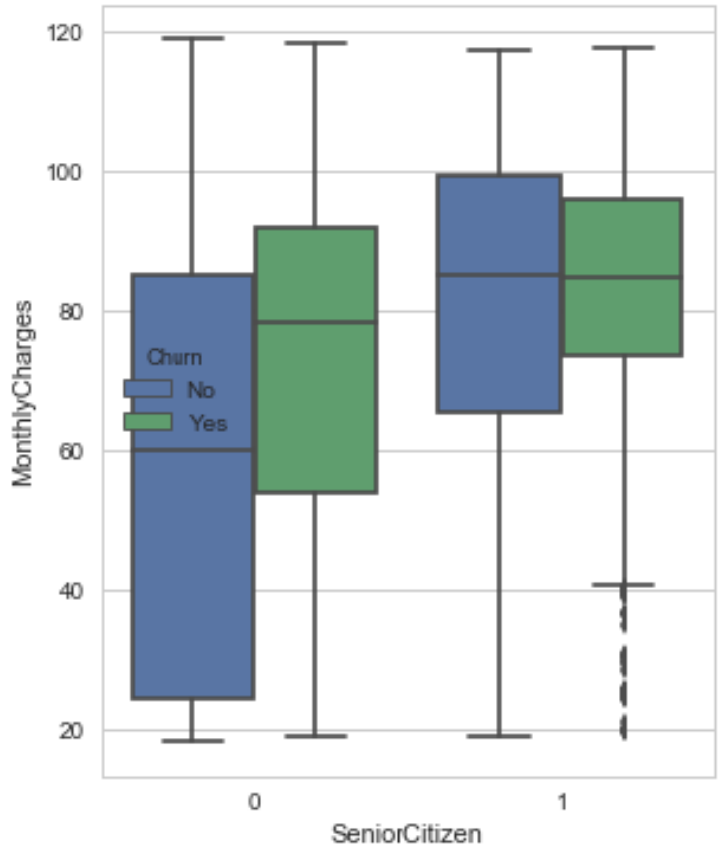
Correlation between continuous variables



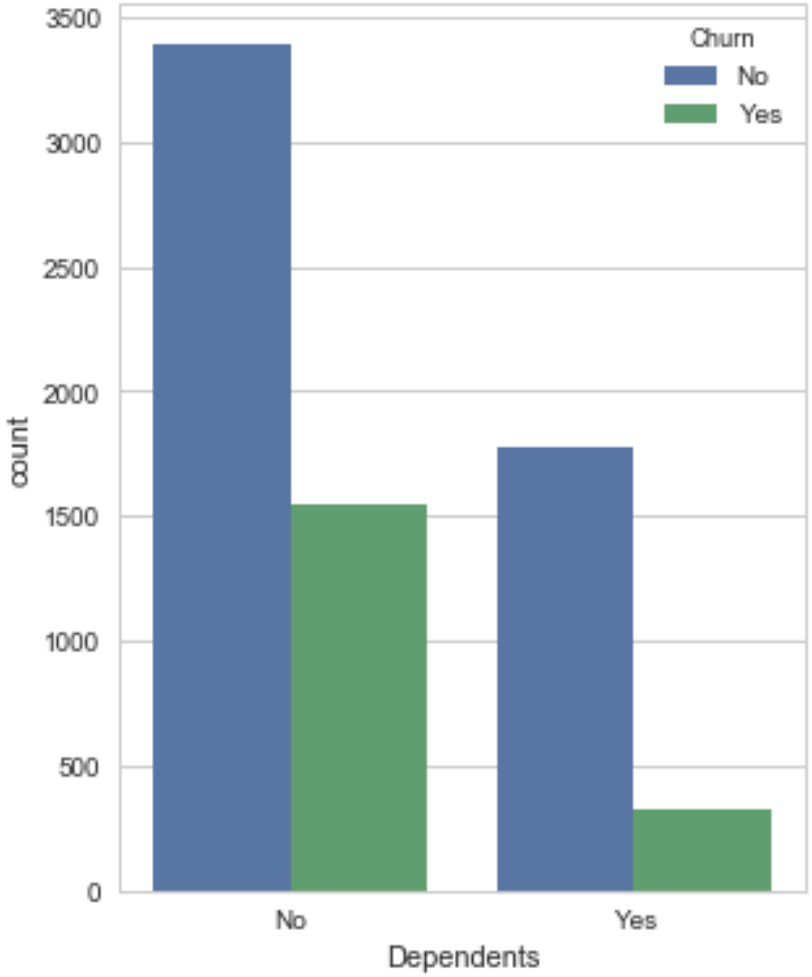
How do senior citizens compare?

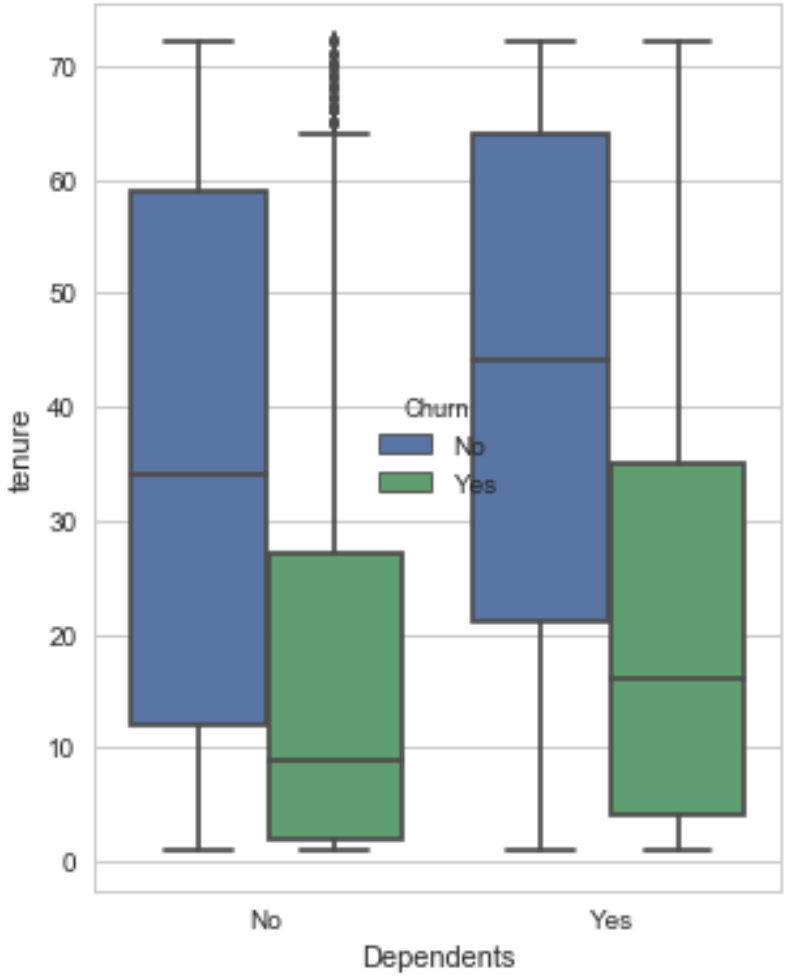


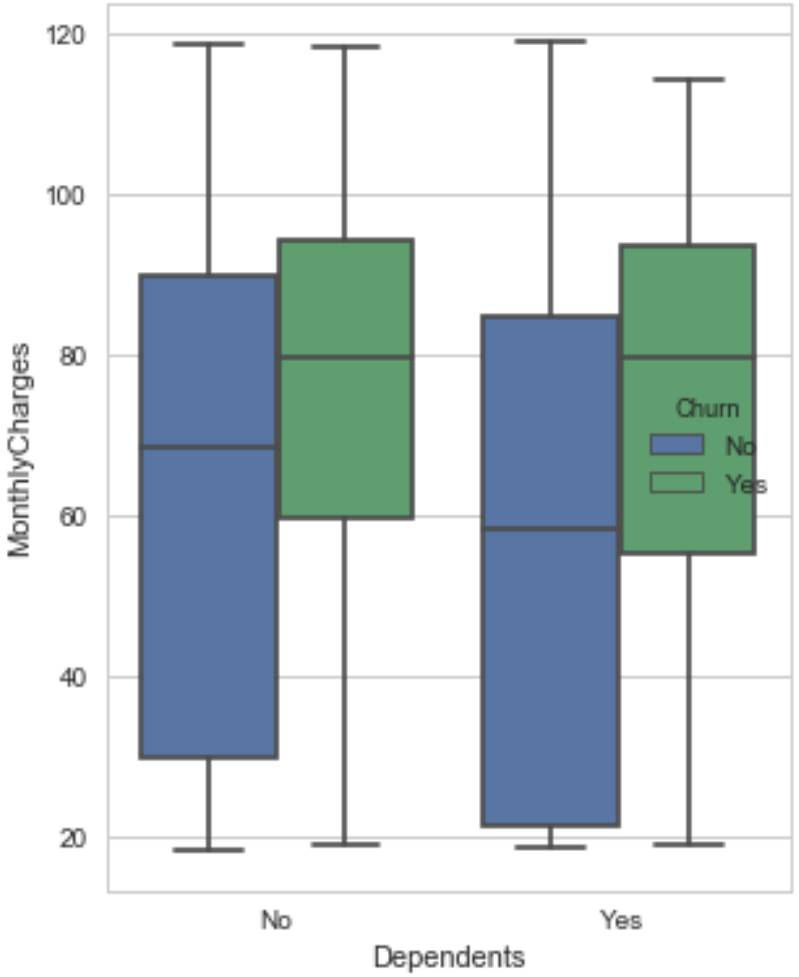




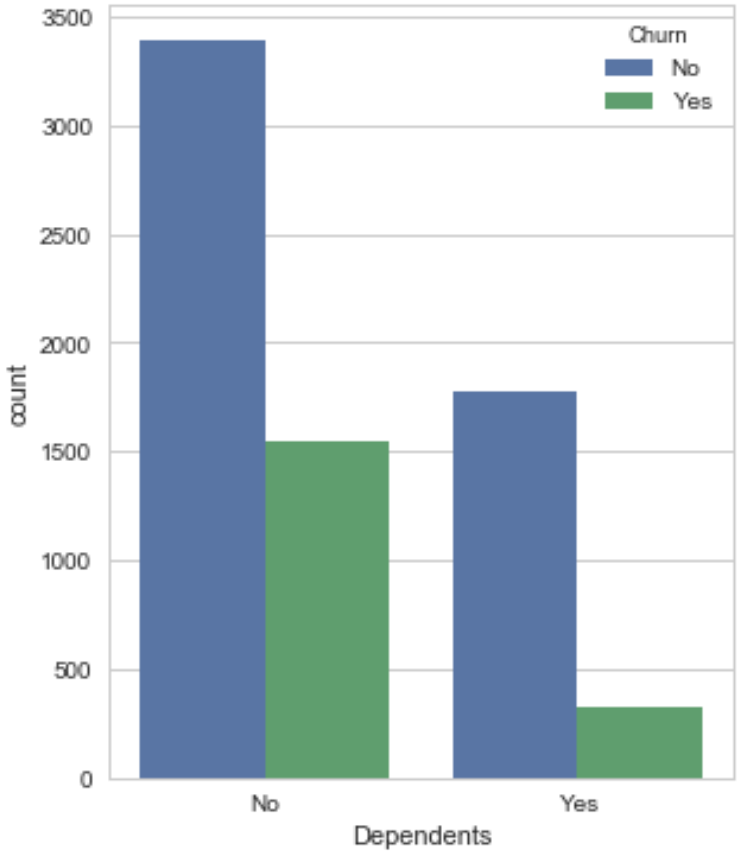
Dependents

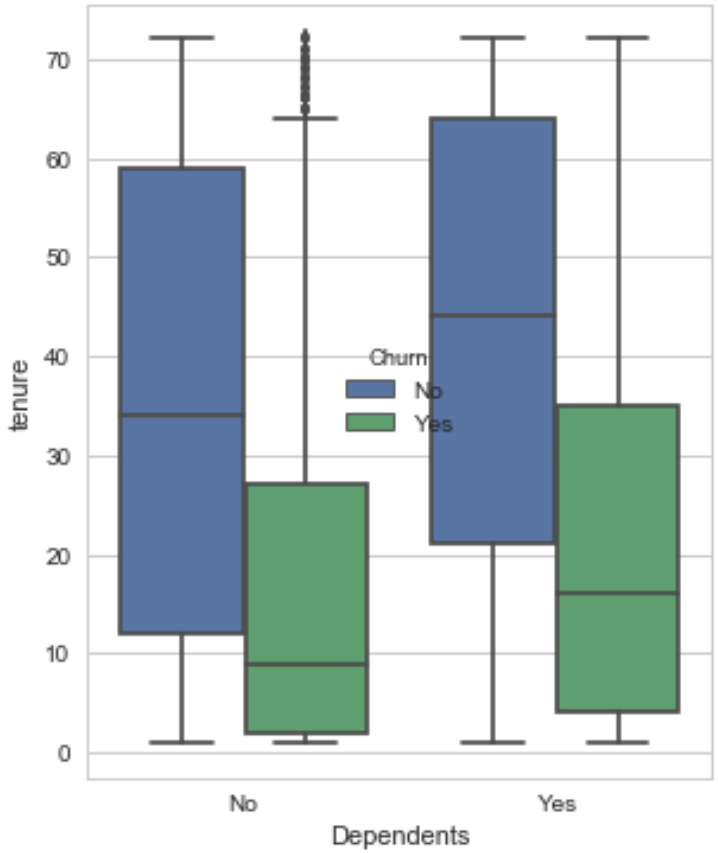

Churn
No
Yes
count
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Dependents

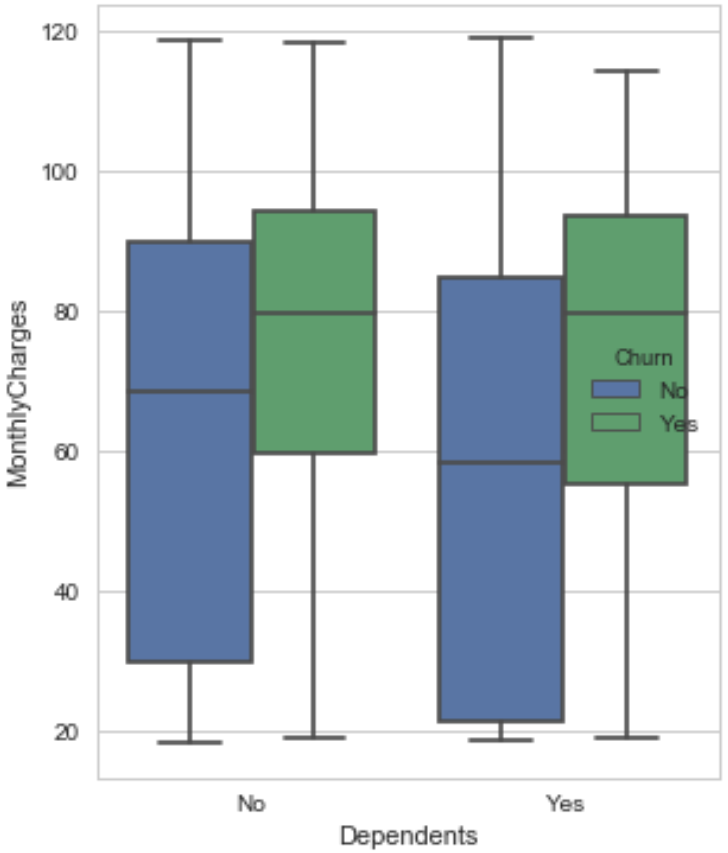

Churn
No
Yes
tenure
70
60
50
40
30
20
10
0
No
Yes
Dependents


Churn
No
Yes
MonthlyCharges
120
100
80
60
40
20
No
Yes
Dependents

Dependents


Churn
No
Yes
count
3500
3000
2500
2000
1500
1000
500
0
No
Yes
Dependents


tenure
70
60
50
40
30
20
10
0
Churn
No
Yes
No
Yes
Dependents


MonthlyCharges
120
100
80
60
40
20
Churn
No
Yes
No
Yes
Dependents

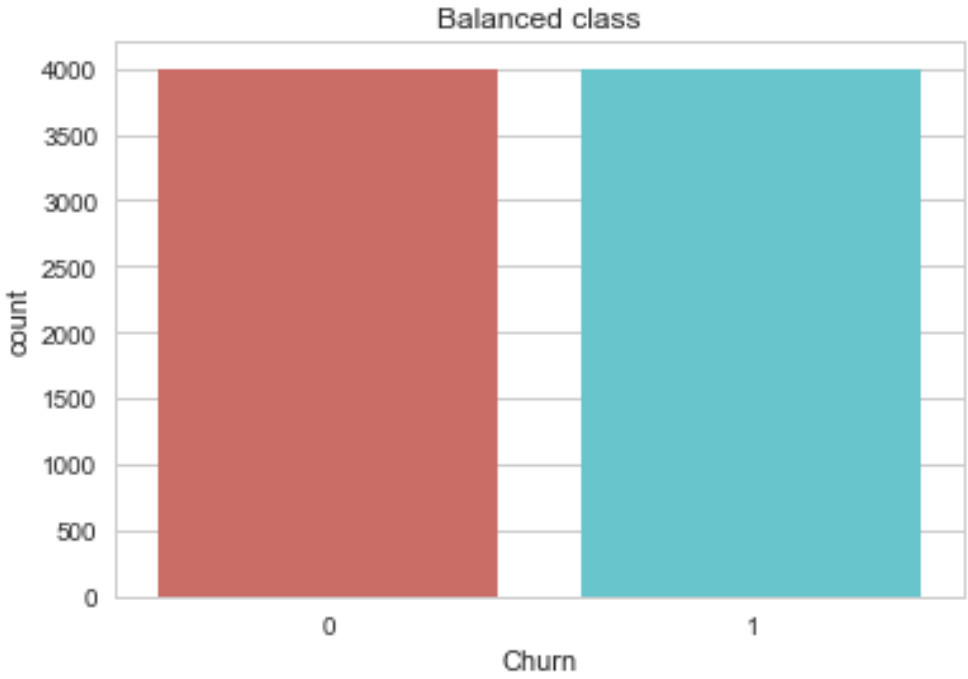
Preparing for Data for Modeling

- Handling categorical variables
- Considerations
- Class imbalance
- Train-Test Split

Class Imbalance


```
0    5163
1    1869
Name: Churn, dtype: int64
```

```
1    4000
0    4000
Name: Churn, dtype: int64
We now have a sample with balanced class
```



Models

- Naive Bayes Bernoulli Classifier
- K-nearest neighbors (KNN)
- Logistics Regression
- Ridge Classifier
- Lasso (Logistics regression with L1 regularization parameter)
- Decision Tree Classifier
- Random Forest
- Support Vector Classifier (SVC)
- Gradient Boost Classification

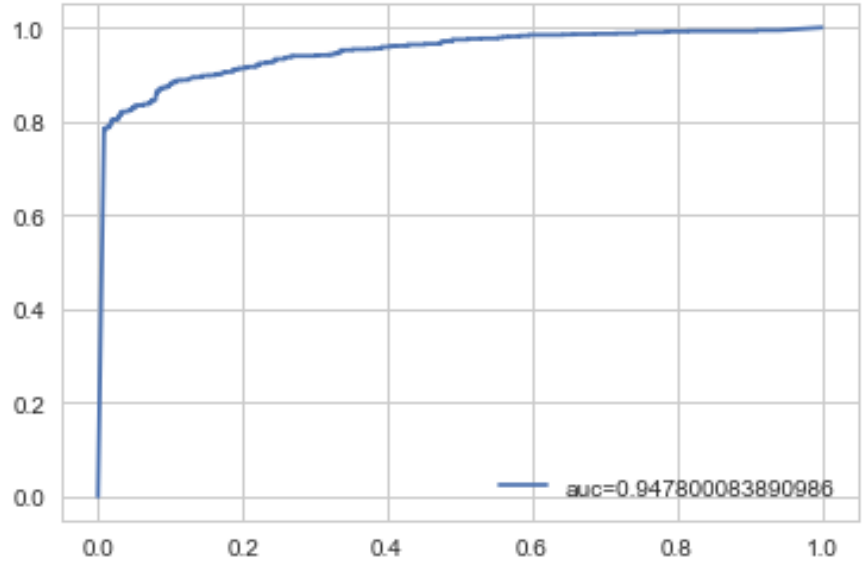
Model Evaluation Methods

- Accuracy score
- Confusion matrix
- Classification report
- AUC (Area Under the Curve)
- Run time

KNN

Classifi. report	Precision	Recall	F1-Score
Churn No (0)	0.92	0.74	0.82
Churn Yes (1)	0.78	0.94	0.85
Avg / total	0.85	0.84	0.84

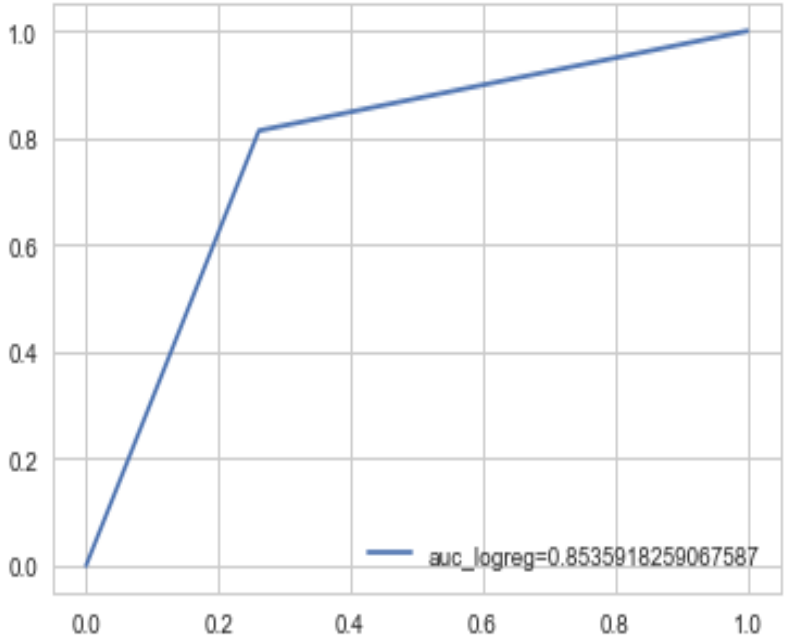
- Cross validation score: [0.84 0.8425 0.8325 0.86375 0.82625 0.83625 0.8325 0.84625 0.84625 0.85375]
- Training set score: 0.9926785714285714
- Confusion matrix
- [[883 311]
  [78 1128]]
- Run time: 0.74 sec
- Auc score: 0.94
-



Logistics Regression

	Precision	Recall	F1-Score
Churn No (0)	0.8	0.74	0.77
Churn Yes (1)	0.76	0.81	0.79
Avg / total	0.78	0.78	0.78

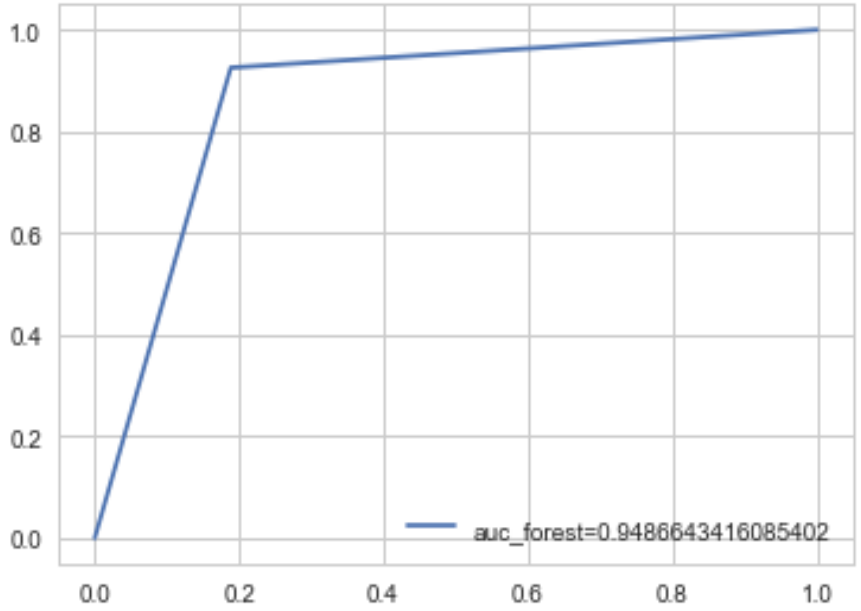
- Training set score: 0.75875
- Test set score: 0.77625
- Cross val score: [0.76125 0.72625 0.75375 0.78375 0.75875 0.775 0.7625 0.745 0.775 0.79125]
- Confusion matrix
- [[882 312]
- [225 981]]
- Run time: 0.91 sec
- AUC score: 0.853



Random Forest

	Precision	Recall	F1-Score
Churn No (0)	0.92	0.81	0.86
Churn Yes (1)	0.83	0.93	0.88
Avg / total	0.87	0.87	0.87

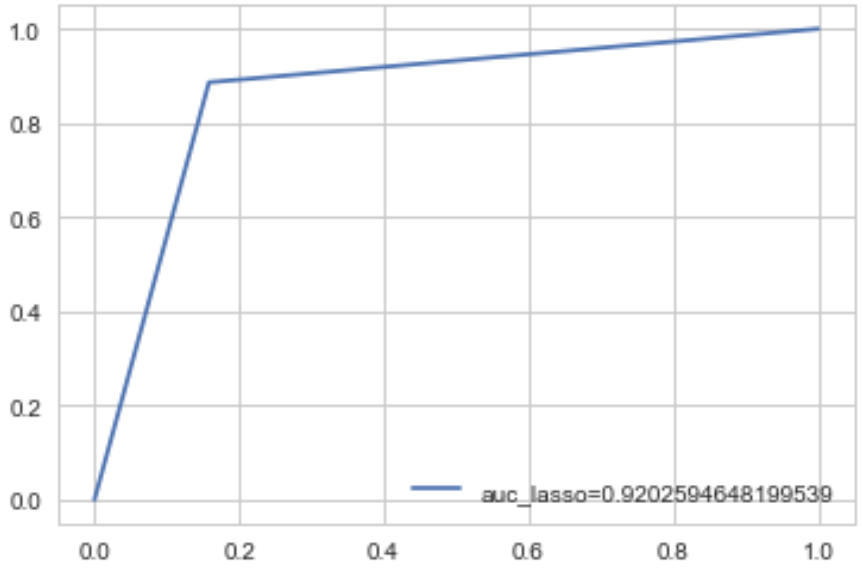
- GridSearchCV params: {'criterion': 'gini', 'max_depth': 30, 'n_estimators': 250}
- Cross validation: [0.89375 0.87875 0.86375 0.91125 0.8675 0.895 0.875 0.89125 0.90375
- 0.89625]
- Confusion matrix
- [[969 225]
- [90 1116]]
- Run time: 65 min
- AUC score: 0.948



SVM

	Precision	Recall	F1-Score
Churn No (0)	0.88	0.84	0.86
Churn Yes (1)	0.85	0.89	0.87
Avg / total	0.86	0.86	0.86

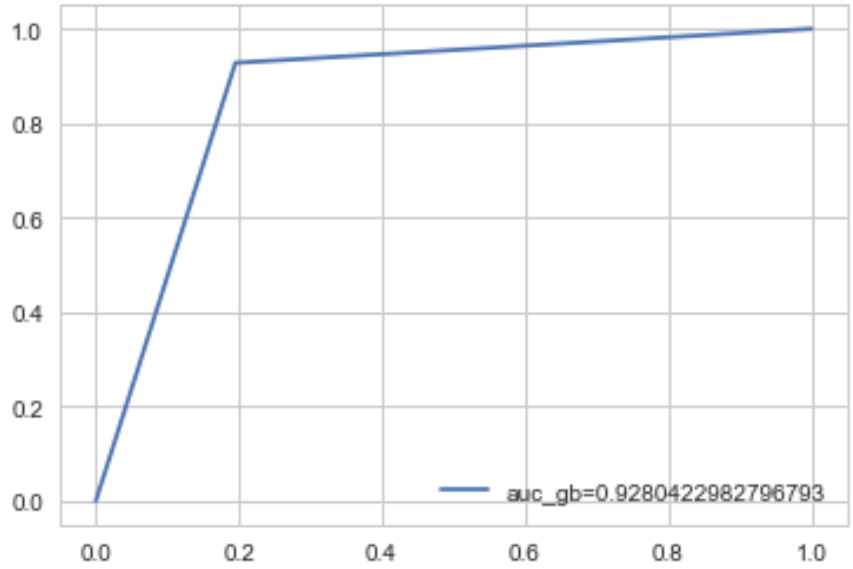
- Cross validation score [0.89375 0.87875 0.86375 0.91125 0.8675 0.895 0.875 0.89125 0.90375
- 0.89625]
- Training set score: 0.9858928571428571
- Test set score: 0.8641666666666666
- Confusion matrix
- [[1005 189]
- [137 1069]]
- Run time 297 mins
- AUC score: 0.920



Gradient Boost

	Precision	Recall	F1-Score
Churn No (0)	0.92	0.80	0.86
Churn Yes (1)	0.83	0.93	0.87
Avg / total	0.87	0.87	0.87

- Cross validation score: [0.92872109 0.92943281 0.91879922 0.91679766 0.94396875]
-
- Confusion matrix
- [[961 233]
- [87 1119]]



- Run time 74 min
- AUC score: 0.928

Overall comparision

Models	Score	Precision	Recall	AUC	Runtime
Naive Bayes Bernoulli Classifier	0.71	0.71	0.71	0.76	0.25 sec
KNN	0.83	0.85	0.84	0.94	0.74 sec
Logistics Regression	0.77	0.78	0.78	0.85	0.91 sec
Ridge Classifier	0.77	0.77	0.77		0.19 sec
Lasso	0.77	0.78	0.78	0.85	3.26 sec
Decision Tree Classifier	0.84	0.85	0.84	0.84	20.01 sec
Random Forest	**0.87**	**0.87**	**0.87**	**0.94**	**65 min**
Support Vector Classifier (SVC)	0.86	0.86	0.86	0.92	5 hrs
Gradient Boost Classification	0.92	0.87	0.87	0.92	75 min

Dimentionality reduction with PCA

- Retain 90% of varience
- 15 components from 44 features
- Resulted in no improvement

KNN with PCA

Classifi. report	Precision	Recall	F1-Score
Churn No (0)	0.92	0.70	0.79
Churn Yes (1)	0.76	0.94	0.84
Avg / total	0.84	0.82	0.82

- Cross validation score:[0.81875 0.84 0.80875 0.84375 0.815 0.8225 0.815 0.84625 0.84
- 0.835]
- Training set score: 0.998035714285714

Confusion matrix

[[830 364]

[69 1137]]

- Run time: 2.98 sec
- Auc score: 0.95
-

Logistics Regression with PCA

	Precision	Recall	F1-Score
Churn No (0)	0.79	0.74	0.76
Churn Yes (1)	0.75	0.80	0.78
Avg / total	0.77	0.77	0.77

- Training set score: 0.7516071428571428
- Test set score: 0.76875
- Cross val score: [0.76 0.72625 0.75 0.7725 0.745 0.76 0.75375 0.73875 0.7725
- 0.79]
- Confusion matrix
- [[879 315]
- [240 966]]
- Run time: 0.35 sec
- AUC score: 0.84

Random Forest PCA

	Precision	Recall	F1-Score
Churn No (0)	0.89	0.82	0.85
Churn Yes (1)	0.83	0.90	0.86
Avg / total	0.86	0.86	0.86

- forest_para: {'criterion': 'gini', 'max_depth': 50, 'n_estimators': 70}
- Cross validation: [0.89125 0.88125 0.855   0.88875
  0.875   0.8875 0.87625 0.88375 0.88875
- 0.87625]
- Confusion matrix
- [[975 219]
- [123 1083]]
- Run time: 2.8 hrs
- AUC score: 0.92

SVM with PCA

	Precision	Recall	F1-Score
Churn No (0)	0.87	0.74	0.80
Churn Yes (1)	0.78	0.89	0.83
Avg / total	0.82	0.82	0.81

- Cross validation score [0.89125 0.88125 0.855 0.88875 0.875 0.8875 0.87625 0.88375 0.88875
- 0.87625]
- Training set score: 0.9335714285714286
- Test set score: 0.8158333333333333
- Confusion matrix
- [[882 312]
- [130 1076]]
- Run time 91 mins
- AUC score: 0.859

Gradient Boost with PCA

	Precision	Recall	F1-Score
Churn No (0)	0.89	0.80	0.85
Churn Yes (1)	0.82	0.90	0.86
Avg / total	0.86	0.85	0.85

- Cross validation score: [0.93680234 0.92357344 0.93288672 0.92189922 0.94178359]
- Confusion matrix
- [[959 235]
- [116 1090]]

- Run time 74 min
- AUC score: 0.924
-

Feature Selection with SelectKBest

- Select 20 best features

KNN with selectKBest

Classifi. report	Precision	Recall	F1-Score
Churn No (0)	0.92	0.74	0.82
Churn Yes (1)	0.78	0.94	0.85
Avg / total	0.85	0.84	0.84

- Cross validation score:[0.8425 0.8375 0.83125 0.85875 0.82   0.83875 0.8325 0.8375 0.83375
- 0.85   ]
- Training set score: 0.9926785714285714

Confusion matrix

[[883 311]
 [78 1128]]

- Run time: 0.65 sec
- Auc score: 0.94
-

Logistics Regression with SelectKBest

	Precision	Recall	F1-Score
Churn No (0)	0.80	0.72	0.76
Churn Yes (1)	0.75	0.82	0.78
Avg / total	0.77	0.77	0.77

- Training set score: 0.7501785714285715
- Test set score: 0.7691666666666667
- Cross val score: [0.755 0.7275 0.74 0.78125 0.76125
  0.7575 0.755 0.74375 0.76375
- 0.77375]
- Confusion matrix
- [[860 334]
- [220 986]]
- Run time: 0.6 sec
- AUC score: 0.84

Random Forest with SelectKBest

	Precision	Recall	F1-Score
Churn No (0)	0.90	0.82	0.85
Churn Yes (1)	0.83	0.91	0.87
Avg / total	0.86	0.86	0.86

- forest_para: {'criterion': 'gini', 'max_depth': 70, 'n_estimators': 400}
- Cross validation: [0.87875 0.86625 0.8625 0.88875 0.85375 0.86875 0.85875 0.8675 0.88
- 0.89]
- Confusion matrix
- [[975 219]
- [114 1092]]

- Run time: 61 min
- AUC score: 0.93

SVM with SelectKBest

	Precision	Recall	F1-Score
Churn No (0)	0.87	0.89	0.88
Churn Yes (1)	0.89	0.86	0.88
Avg / total	0.88	0.88	0.88

- Cross validation score [0.89125 0.88125 0.855 0.88875 0.875 0.8875 0.87625 0.88375 0.88875
- 0.87625]
- Training set score: 0.9810714285714286
- Test set score: 0.8766666666666667
- Confusion matrix
- [[1063 131]
- [165 1041]]
- Run time 5 hrs 30 mins
- AUC score: 0.918

Gradient Boost with SelectKBest

	Precision	Recall	F1-Score
Churn No (0)	0.89	0.80	0.85
Churn Yes (1)	0.82	0.90	0.86
Avg / total	0.86	0.85	0.85

- Cross validation score: [0.91421875 0.91798281 0.90358828 0.8971125 0.93073906]
- Confusion matrix
- [[949 245]
- [119 1087]]
- Run time 32 min
- AUC score: 0.91
-

Overall comparision based on CV score

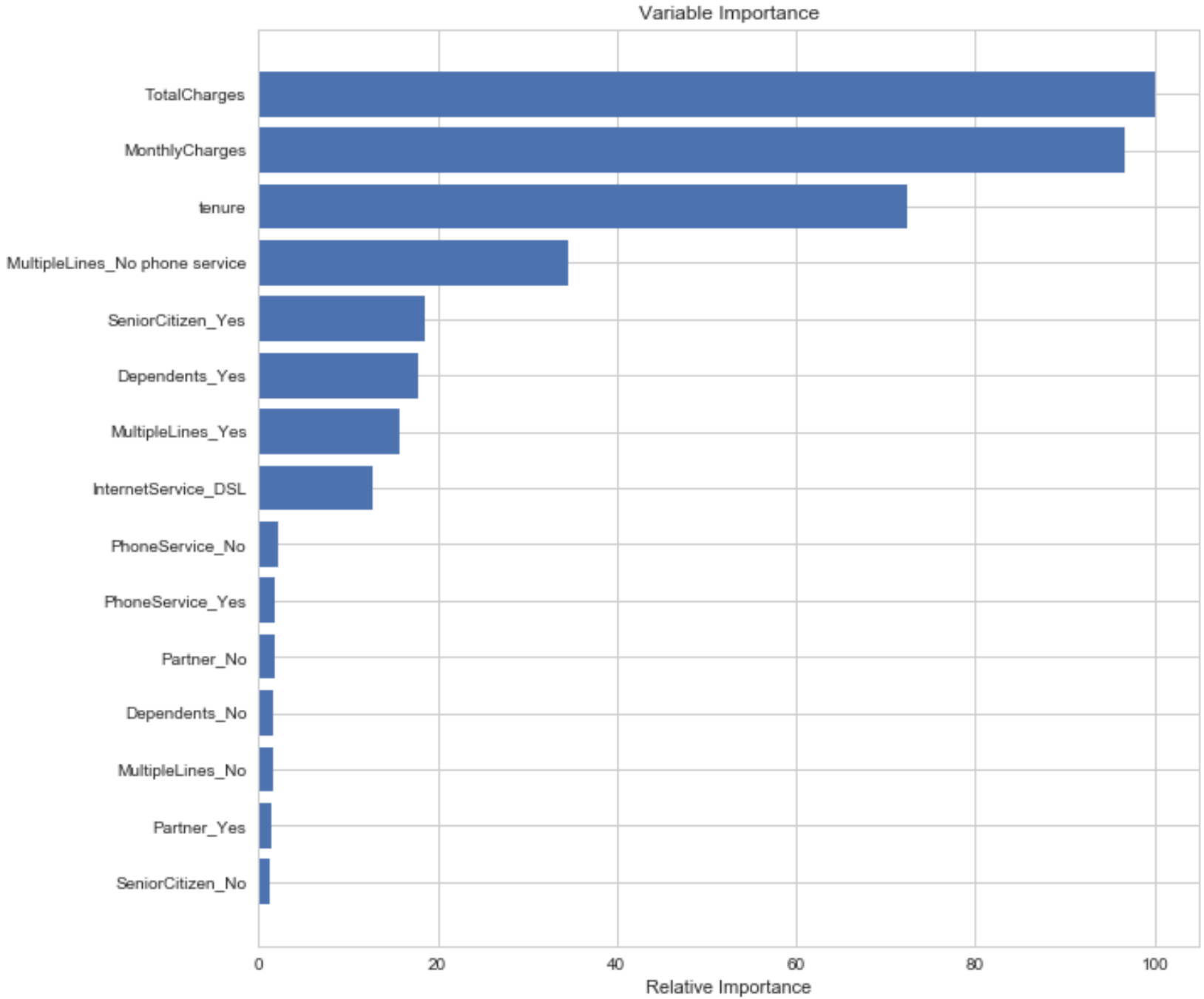
Models	**Original Features**	**PCA**	**SelectKBest**
KNN	0.83	0.81	0.83
Logistics Regression	0.77	0.76	0.76
Random Forest	**0.87**	**0.88**	**0.87**
Support Vector Classifier (SVC)	0.86	0.81	0.87
Gradient Boost Classification	0.92	0.93	0.91

Recommended Model

- Random Forest
  - Score
  - Runtime
  - Recall and precision

Feature Importance

1. Total Charge
2. Monthly Charge
3. Tenure
4. Whether they have phone service or not
5. Whether the customer is senior citezen


Variable Importance
TotalCharges
MonthlyCharges
tenure
MultipleLines_No phone service
SeniorCitizen_Yes
Dependents_Yes
MultipleLines_Yes
InternetService_DSL
PhoneService_No
PhoneService_Yes
Partner_No
Dependents_No
MultipleLines_No
Partner_Yes
SeniorCitizen_No
0
20
40
60
80
100
Relative Importance

Practical use of the Model

- Identify which customer group to focus on
- See if hidden factors are causing churn and fix them
- Optimize service offering

Conclusion and Recommendation

- Customers are leaving mostly because of charge
  - Check if competition is offerng better price
  - Look into the business model to find ways to reduce charge
  - Focus on seniors, customers in their first months of tenure
  - Provide family focused service for people with dependents

Limitations of the Model

- Size of data
- Loss and duplication of information
- More parameter tuning