CS7641 - Supervised Learning Report

Shawn R. Mailo

September 24, 2018

Data Sets

- Cancer Data set

This data set is called the Wisconsin Breast Cancer data set and comes from SciKit Learn's Sample Data sets. This Data set is a classic binary classification data set in which the target attribute is a diagnosis if the patient has breast cancer or not. This data set contains 569 records of separate/individual patients. The dimension of this data set is 30 and all non-target attributes are real, positive, continuous numerical data types. The non-target attributes are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image and include [mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, compactness error, concavity error, concave points error, symmetry error, fractal dimension error, worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst compactness, worst concavity, worst concave points, worst symmetry, worst fractal dimension]. This data set is slightly skewed toward malignant classifications at 63%. This data is interesting because all features are continuous numerical data types and contains a limited amount of training data both of which differentiate this data set from the other data set of this analysis.

- Churn Data set - Preproceesing

In theory, when a model has more attributes, the training set needs more data to maintain accuracy. Since this data set has a relatively small amount of training data some attributes were dropped. These attributes included all the error attributes (as it was assumed that their relationship to the target attribute is less/non significant). After creating swarm, violin, and correlation plots, if an attribute showed little to no correlation to the target output or high correlation with another non-target attribute then this attribute was dropped. After feature selection, the original 30 attributes now only consisted of 8. The data set was then split into a training set and a test set. 80% of the data set was allocated to the training set while the other 20% was set aside as the test set. Lastly all the data was scaled to make the distributions standardized. The scaling factors were applied to both training and test sets, however the scaling factors were only fitted to the training set as to not incur bias from the testing set.

- Churn Data set

This data set comes from IBM's Sample Data sets and is used to predict the customer behavior in order to develop focused customer retention programs. This data set has 7043 rows of data.

Every row in the data set represents a customer while the columns contains the 21 attributes of the customer. The attributes of the customers include [CustomerID, Gender,SeniorCitizen, HasPartner,HasDependents, Tenure, PhoneService, MultipleLines, InternetServiceProvider, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, ContractType, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn]. All attributes are categorical except for Tenure, MonthlyChares, and TotalCharges which are numerical. The target attribute is Churn which specifies if the customer discontinues their subscription to the service. Since Churn, the target attribute, is only either 'Yes' or 'No', all models that will be created will be considered binomial classification models. The data set is skewed with 73% of customers having 'No' for their churn attribute. This data is interesting because majority of the attributes are categorical.

- Churn Data set - Preproceesing

In order to create predictive models, some preprocessing steps were implemented in aims to improve accuracy and speed. First the Totalcharges attribute was converted from a string object to a numeric object. 11 rows of data were then drop due to having NULL values. Since CustomerID should have no relationship with Churn rates, this column was dropped from the data set. Some categorical attributes had multiple types of classes which could lead to lower model performance. To address this, these attributes were converted into dummy/binary indicator variables. The next step was to split the data into training and testing sets. 80% of the data set was allocated to the training set while the other 20% was set aside as the test set. Lastly all the data was scaled to 0,1 to make distributions normalized. The scaling factors were applied to both training and test sets, however the scaling factors were only fitted to the training set as to not incur bias from the testing set.

Decision Tree

Decision Tree - Learning Curve

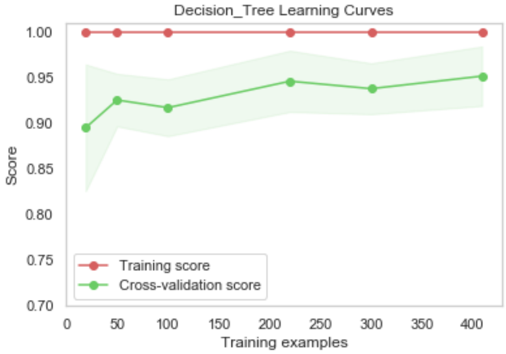

Figure 1: Cancer Data Set

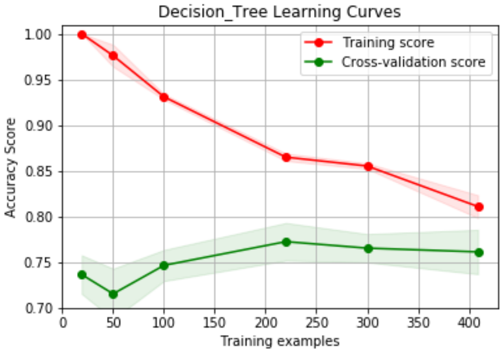

Figure 2: Churn Data Set

From the learning curve of the Cancer Data Set it can be seen that when comparing the training and CV score curves that both become steady state with training sets greater than 250 examples.

They also have a consistent and significant gap. Due to this steady state gap and that the training accuracy is at 100%, it can be assumed that the model is over fitting and there is high variance.

The learning curve of the Churn Data Set shows the descending training accuracy converging into the CV accuracy. Since the two lines are converging and not have become stagnant, it can be assumed that more training samples would allow the lines to continue to converge, leading to a better model. Since it seems that the training curve is converging to close the gap with the CV curve, and there is room for improvement, it can be assumed that this model has low variance and still high bias and that a more complex algorithm or more complex data features would help improve accuracy.

Decision Tree - Complexity Curve

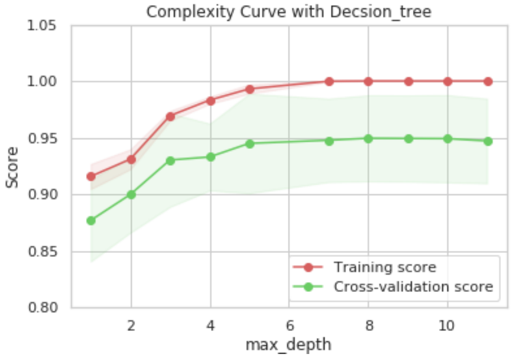

Figure 3: Cancer Data Set

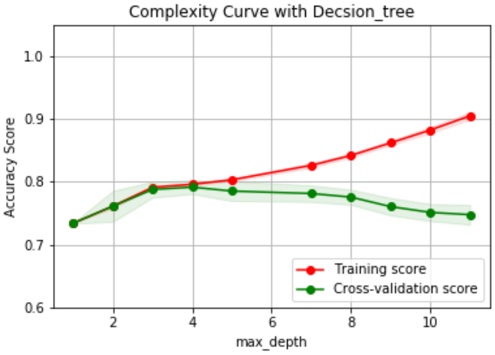

Figure 4: Churn Data Set

These complexity curve graphs show training scores consistently improving and even hitting 100% in the Cancer Data Set. Meanwhile the CV scores increase but start to decline once max-depth goes above a certain value. This means that there a sweet spot where the value of max-depth allows for optimum scores on test sets and going below this number would signify under fitting while going higher would signify over fitting. In the Churn Data Set, if max-depth continued to increase (increasing the over fitting) the training score would most likely continue to increase while the CV would continue to decrease.

Boosting

Boosting - Learning Curve

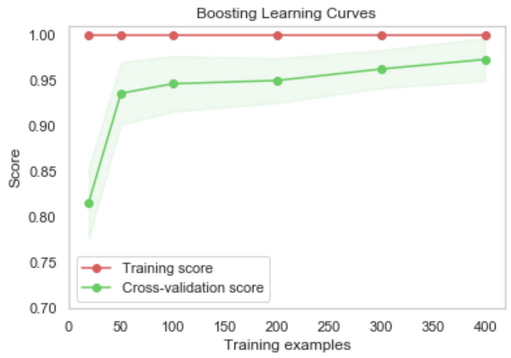

Figure 5: Cancer Data Set

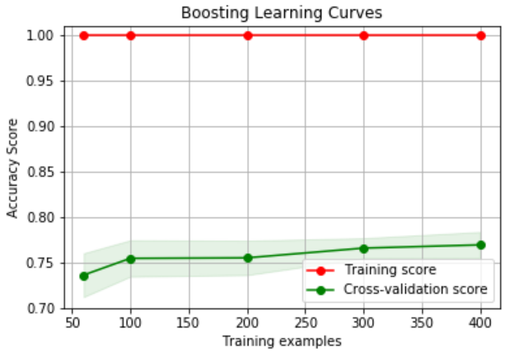

Figure 6: Churn Data Set

Cancer Data Set - Training curve stagnates close to 100% accuracy, which possibly means over fitting. The The CV score curve starts with a big gap between the training score curve but converges towards the training score curve. If there were more samples this gap would continue to decrease. A small gap signifies low variance. Due to this and that fact that both curves are converging to 100% accuracy it can be assumed the model is very accurate and has minimized the effects of over and under fitting.

Churn Data Set - Training curve is stagnant around 100% while the CV score curve is slowly increasing as the number of training samples increase. Additionally there is a big gap between the two lines which signifies high variance and over fitting. More samples maybe needed to allow the CV score curve to continue to rise but the model most likely needs to be more regularized in order to combat over fitting.

Boosting - Complexity Curve

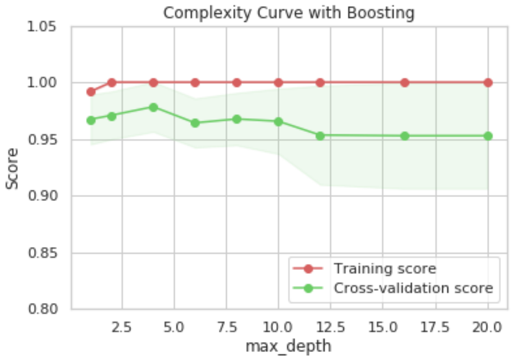

Figure 7: Cancer Data Set

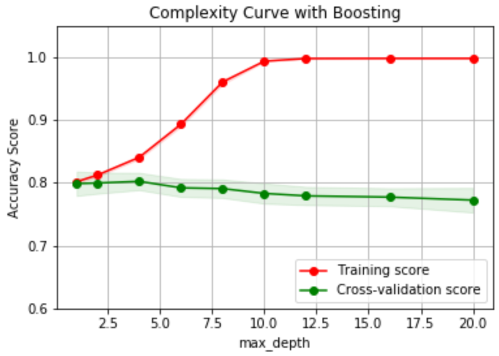

Figure 8: Churn Data Set

Similar to the decision tree, here the hyper parameter 'max-depth' of the trees used in the boost models was tuned to maximized accuracy.
Cancer Data Set - Here the max-depth shows an optimum value of about 4. The CV score curves increases until 4 and begins to decrease afterward while training curves stays at about 100%. Below

4 is under fitting while above 4 is over fitting.
Churn Data Set - Here the max-depth shows an optimum value of about 4. After 4 the model begins to over fit the training data which can be seen as training data goes to about 100%. Due to the large gap between the training score and CV score curve at the optimized max-depth point, it can be assumed that there is high variance and that regularizing other parameters may help the accuracy of the model.

Neural Networks

- Learning Curve

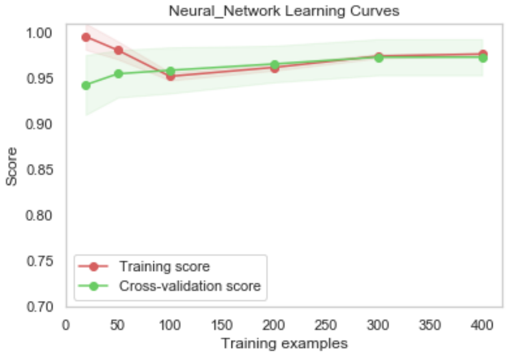

Figure 9: Cancer Data Set

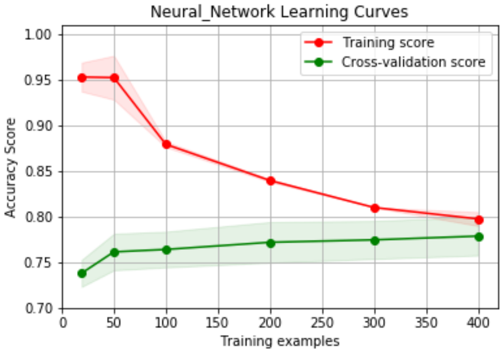

Figure 10: Churn Data Set

Both graphs show the training accuracy starting off high and converging down to the CV score curve. This shows that Neural Networks tend to over fit with low numbers of training samples and regularize/generalize more when given larger sets of training examples. Cancer Data Set - Both curves converge together leaving a small gap, both having a high accuracy. This suggests a high accurate model with minimal bias and variance.

Churn Data Set - Both curves converge together leaving a small gap at about 80 % accuracy. The small gap signifies low variance, however to 20% inaccuracy signifies a bias. This bias can be decreased by using a more complex model or more complex data features. An example of increasing the complexity of a Neural Network is to increase the number of hidden layers or increasing the number of nodes at each hidden layer.

Neural Network - Iteration Learning Curve

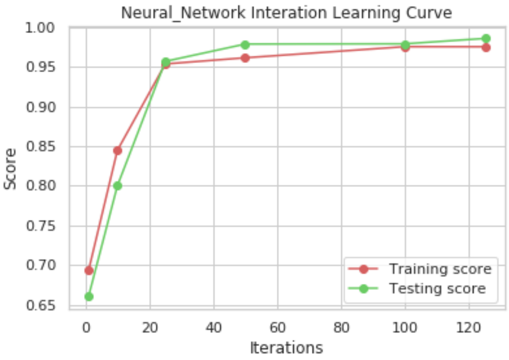

Figure 11: Cancer Data Set

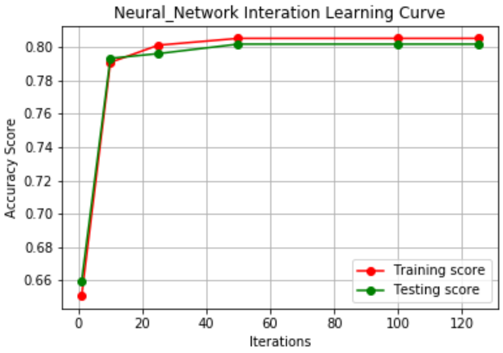

Figure 12: Churn Data Set

For Neural Networks the model tries to fit the training data by doing gradient descent calculations in iterations. Each iteration to get the model to a local minimum of error. Theoretically the more iterations performed the more accurate the model will be. This can be seen on both Data set models. The score curves increase greatly from 1 iteration to 20 iterations. After this the improvements become minimal. However, increasing iterations do not make up for model bias which can be seen in the Churn data set model. There is very little gap between the score curves but both max out at around 80% accuracy. This suggests low variance and high bias. To address this, one can tune hyper parameters, use more complex models, add more features.

- Complexity Curve

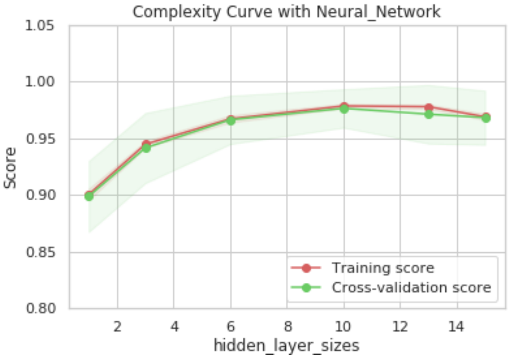

Figure 13: Cancer Data Set

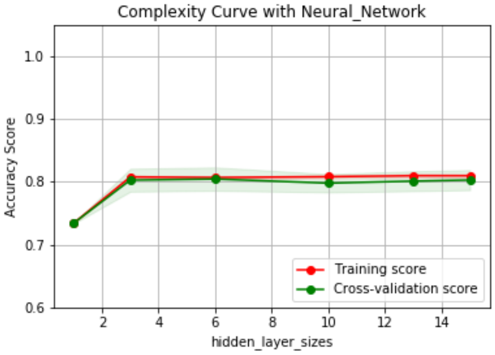

Figure 14: Churn Data Set

For the Neural Network complexity graphs, the number of hidden layers were tuned. Hidden layer sizes ranges from 1-15. Both graphs show a minimal/almost non-existent gap between the training score and the CV score curves. This shows low variance for both models.
Cancer Data Set - The optimal hidden layer size for this data set is at 10. The accuracy score is also very high for both curves, which shows low bias. Over all this model represents the feature-target output relationship well.
Churn Data Set- The optimal hidden layer size for this neural network is around 3. After this the accuracy doesn't improve above 80% for both training and CV score curves. This constant inaccuracy means that the model needs to be more complex or have more data features.

Support Vector Machines

Support Vector Machines - Learning Curve

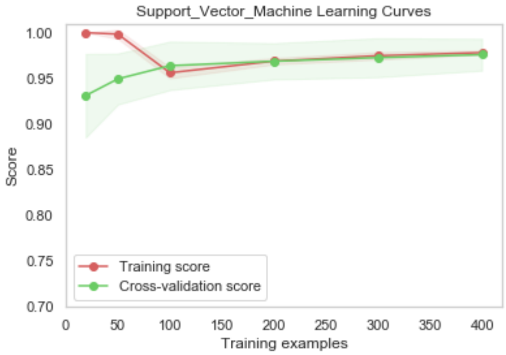

Figure 15: Cancer Data Set

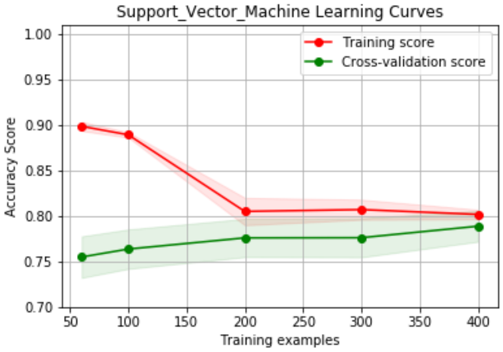

Figure 16: Churn Data Set

It can bee seen in both data sets that the training score curve starts with a high accuracy and decreases and the two curves converge as number of training samples increase. This means that the model tends to over fit with a low number of training samples and generalizes better as number of training samples increase. The small gap in between both curves signifies low variance in the models.

Cancer Data Set - Having low variance and high accuracy for both curves signifies that this model does a good job at representing the relationship between the features and target output.

Churn Data Set - The two score curves converge around 80%. A small gap between the curves with an inaccuracy of 20% means the model has a bias problem and can improve with a more complex model, hyper parameter tuning, and/or more features.

Support Vector Machines - Iteration Learning Curve

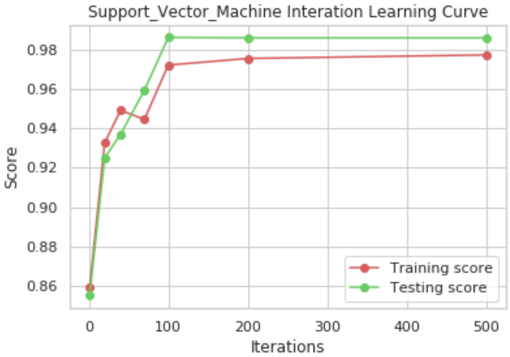

Figure 17: Cancer Data Set

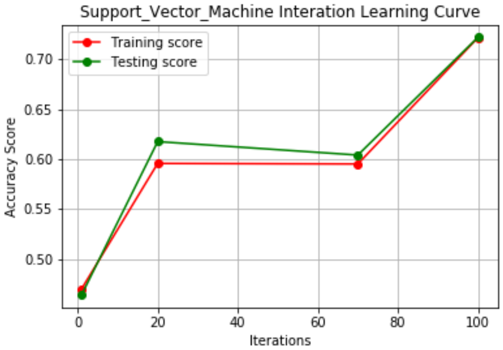

Figure 18: Churn Data Set

Increasing the iterations of the Support Vector Machine is similar to that of the Neural Network. With every iteration it will continue to perfect the separation planes between classes. As iterations increase the accuracy score increases rapidly and tapers out. With each iteration the improvement is a little smaller. However iterations cannot overcome all bias in model. Sometimes the the SVM needs a more complex/representative kernel, or more data features.

Support Vector Machines - Complexity Curve

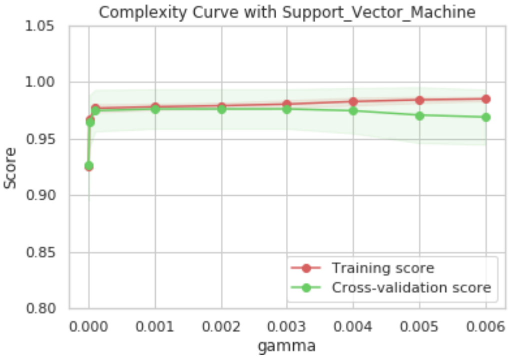

Figure 19: Cancer Data Set

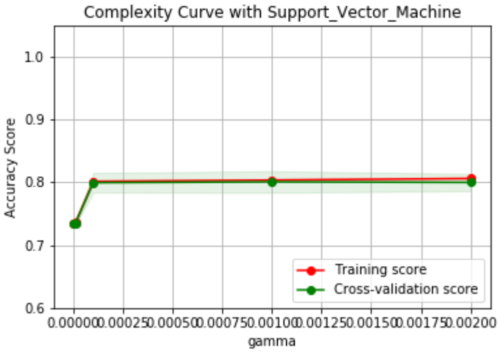

Figure 20: Churn Data Set

Both Data Sets show a similar pattern where the gap between training and CV curves are very small and increase in size as gamma is increased. The gap increases because as gamma increases the training score increases slightly while the CV curve starts to fall. This signifies that lower values of gamma around 0.0001 are optimal for these models.

Cancer Data Set - Having low variance and high accuracy for both curves signifies that this model does a good job at representing the relationship between the features and target output.

Churn Data Set - The optimal gamma brings about 80% accuracy for both the training and the CV score curves. A small gap between the curves with an inaccuracy of 20% means the model has a bias problem and can improve with a more complex model, hyper parameter tuning, and/or more features.

K-Nearest Neighbors

K-Nearest Neighbors - Learning Curve

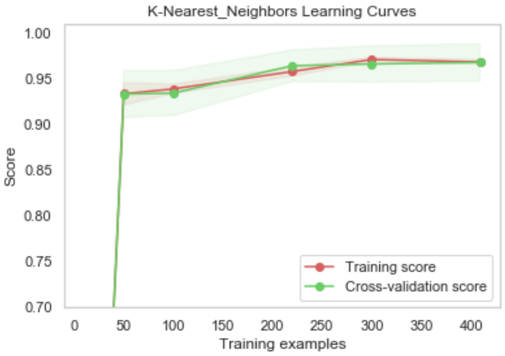

Figure 21: Cancer Data Set

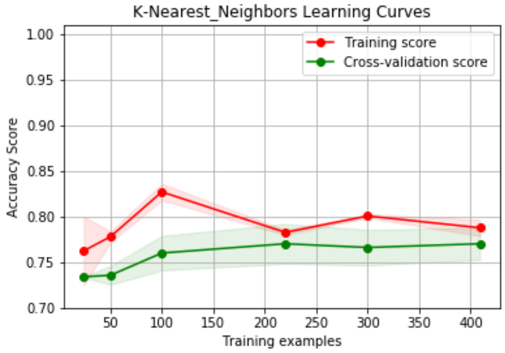

Figure 22: Churn Data Set

*Both models have Number of Neighbors greater than 10

Cancer Data Set - Both score curves performed poorly with a low number of training sets. This shows that KNN models under fits with low number of training examples. Since KNN compares a data point to all of the previous data points, its always better to have more previous data points.

This gives the model a more vast and granular output. This improvement is seen as both curves increase as number of training examples increase. The small gap between the two curves and the high accuracy shows low variance as well as low bias. This model does a good job at representing the relationship between the features and target output.

Churn Data Set - This data set shows both score curves generally increasing as the number of training samples increase. Both score curves stay between 74 and 84% and the training score curve tends to be a few percentages higher than the CV score curve. From this graph it can be said that this model has bias and some variance problems. The bias problem can be addressed by using a more complex kernel to represent distances in a more accurate way or by new adding new features.

K-Nearest Neighbors - Complexity Curve

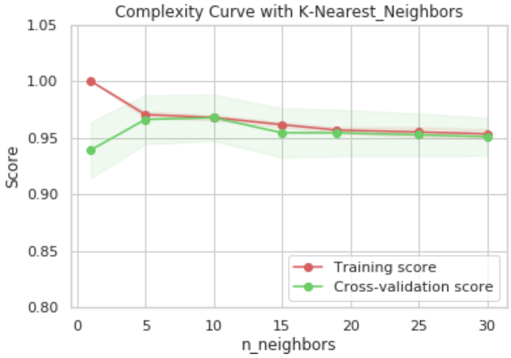

Figure 23: Cancer Data Set

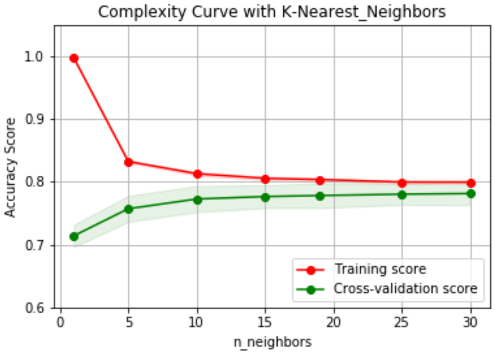

Figure 24: Churn Data Set

Both data sets have a KNN model in which training accuracy starts off at 100% and converges down to the CV score curve. This happens because KNN with number of neighbors = 1 always means 100% accuracy for the train set because each point will match with itself and only itself in the model producing all correct predictions.

Cancer Data Set - Both curves converge between 5-10 number of neighbors. This is also where the CV score curve has a maximum value of about 97%. After this both curves begin the decrease in accuracy at a similar rate. As the number of neighbors parameter increase the more likely the model will under fit because your averaging the classification of more and more data points around the prediction point. Overall with high accuracy of 97% this model has low bias and low variance.

Churn Data Set - The score curves converge at a higher number of neighbors parameter. Since both curves are still converging and have a little gap, the optimum number for number of neighbors parameter may be more than 30. With converged score lines and an inaccuracy of about 20%, this signifies low variance and high bias. As mentioned before, the bias of the model can be improved with a better/ more representative kernel or more features.

Computation Times and Final Test Accuracy Score

Table 1: Cancer Data Set - Computation Times and Final Test Accuracy Score

Model Type	Model fit time (s)	Model Prediction time (s)	Test Score
Decision Tree	0.0038	0.0004	93.2%
Boosting	0.0970	0.0005	96.6%
Neural Network	0.2470	0.0002	99.3%
S.V.M.	0.0041	0.0005	98.6%
K-NN	0.0012	0.0019	95.2%

For the Cancer data set the K-NN model had the fastest fitting time, however it had the slowest prediction time. This is because the KNN model is a Lazy Learner. Performing the number crunching during prediction rather than the fitting process. The Neural Network had the slowest fitting time but the fastest prediction time. This is because the Neural network is an iterative model where computation increases with every iteration. After fitting, the Neural Network model a set of coefficients that are used to perform predictions. This is why the prediction time is quick. SVM and Boosting will have longer fitting times compared to decision trees beacuse SVM is iterable like Neural Networks and Boosting uses many decision trees to make a model.

The final test accuracy scores shows the the neural network was the most accurate with a score of 99.3%. The lowest performer was the decision tree at 93.2% which is still really good for a model. This tells us the this data set isn't a hard data set to make predictions. and that the features have a strong relationship to the target output. I believe the the Neural Network had a higher score because the data was numerical, continuous, and the model was able to be more complex.

Table 2: Churn Data Set - Computation Times and Final Test Accuracy Score

Model Type	Model fit time (s)	Model Prediction time (s)	Test Score
Decision Tree	0.0129	0.0006	79.0%
Boosting	1.8201	0.0046	81.0%
Neural Network	0.4582	0.0008	80.0%
S.V.M.	1.5371	0.2668	79.5%
K-NN	0.0350	0.5801	79.2%

For the Churn Data Set, the Decision tree had the fastest Fitting and Prediction time. Decision trees are pretty simple and fast compared to other models, especially the iterative models such as Neural Networks and SVM. As a lazy learner, KNN had a pretty quick fitting time but had the slowest prediction time. Larger number of samples have greater effects on fitting times for models such as Boosting (because of number of decision trees it needs to make) and Neural Networks and SVM (Number of iterations it has to perform). This effect can be seen when comparing the two different data sets. The Cancer Data Set had a magnitude lower difference in the number of samples which correlates to a much lower fitting time compared to the Churn Data Set.

All models had similar accuracy scores, ranging from 79-81%. This is probably because the models are not complex enough and probably because there are other unknown features that affect the target output. Here the boosting model was the most accurate. One reason that could explain this is that many of the features of this data set are discrete and the boosting model is based on splitting data discretely.