James Kulp

DS 3001 KNN assignment Question 0

1. What is the difference between regression and classification?
    - The main difference between regression and classification are their outputs and the data they are best deployed with. Regression predicts a continuous, numerical value while the classification output is a categorical value, typically binary. Regression uses variables of a dataset to predict a target variable like price, categorical outcomes might be evaluating data and classifying is as spam or not spam like in the case of PCA.
2. What is a confusion table? What does it help us understand about a model's performance?
    - A confusion table helps compare your model's predicted classification to the actual data; they have true positives and negatives and false positives and negatives. It helps quantify how accurate your model is based on the proportion of true to false values along with specificity (how well it avoids false negatives), precision (how well it avoids false positives), among other useful information that give good insight into how well a categorical model is performing.
3. What does the SSE quantify about a particular model?
    - It basically quantifies the difference between all of the predicted values of a model and the actual true values in the data, with any given model you're typically seeking to minimize this value to ensure you have the most accurate model possible.
4. What are overfitting and underfitting?
    - Overfitting would be the case where the model is too complex in that it takes into account too many variables which might be irrelevant to the predicted value and make it less accurate than using less variables but ones which are more directly related to the target variable and don't have much collinearity with one another, it becomes almost too sensitive. Underfitting is the opposite phenomenon when the model is too simple to capture meaningful variation in the data.
5. Why does splitting the data into training and testing sets, and choosing k by evaluating accuracy or SSE on the test set, improve model performance?
    - The Training set allows the model to learn patterns in the data and refine the model while the testing set is useful to evaluate the model's performance on data it has not dealt with yet to ensure that the model refined on the training data is truly as good as it can be. Choosing the best K based on the SSE (accuracy) of the test set does a lot to improve model performance and find the best middle ground between underfitting and overfitting issues.
6. With classification, we can report a class label as a prediction or a probability distribution over class labels. Please explain the strengths and weaknesses of each approach.

- With class labels they are the simplest form of classification which makes it easy for an audience to understand when evaluating the model and looking for meaningful classification outcomes that will be used in decision making. The drawbacks are the classification labels do not contain any uncertainty information which risks misinterpretation or showing how close the model was to identifying it as another label. The advantages of the probability distribution, classification model is that is includes uncertainty information and just how close the model is to making another classification, and might allow a careful user to have more accurate decision making, and allows for more specific threshold requirements if those are appropriate, the drawbacks come from more difficult interpretability, a more complicated model structure which could be overfitting, and an over exaggerated sense of certainty.