QUESTION Should we move the loading of libraries all at the top? Or should we only load them when we’re ready to use them?

**DATA PREPARATION**

[10 points]

done - Define and prepare your class variables.

done - Use proper variable representations (int, float, one-hot, etc.).

done - Use pre-processing methods (as needed) for dimensionality reduction, scaling, etc.

done - Remove variables that are not needed/useful for the analysis.

**Data Selection**

QUESTION Should we change CLASSIFICATION/REGRESSION 1 to Crime Solved or Unsolved Prediction. We did use another title in section 1.1.3.

RECOMMENDATION Change description

The first classification model will be predicting whether a crime will be solved or go unsolved using the variables collected during the initial recording of the homicide data when it was filed.

QUESTION Should we change CLASSIFICATION/REGRESSION 2 to Homicide Perpetrators Prediction. We did use another title in section 1.1.4.

RECOMMENDATION Change description

The second classification model will predict the demographics of a homicide perpetrator (age, race, sex, and ethnicity) using the variables collected during the initial recording of the homicide data when it was filed. Predicting a perpetrators' demographics may result in leads which causes a case to be solved. Sensitivity towards police bias, ethics of profiling, and those falsely incarcerated are a high priority. While predicting who the perpetrator may be can lead to an increased rate of solved cases. Careful consideration will be given to avoid wrongfully accusing a suspect.

RECOMMENDATION Remove

from itertools import chain

from geopy.geocoders import Nominatim

**DATA PREPARATION**

**subsection 1.1**

QUESTION We should decide if we want the full write-up at the top of the section or inside of the subsections 1.1.x. I like a quick summary at the top, then the more robust summary in each section 1.1.x

RECOMMENDATION This sentence should be changed if its a part of the numbered list

While the age groups were binned, the original ages are \*\*not\*\* necessary for model analysis, hence, they're removed for these purposes.

RECOMMENDATION Update sentence

Object data types are typecasted to categorical.

RECOMMENDATION Update sentence

Two dataframes are generated for each classification model.

QUESTION Should we use the name of the classifications versus classification 1 & 2

Although, both use the same attributes, with the addition of concatenated perpetrator data for classification 2, only the dataframe for classification 1 is dummy coded.

RECOMMENDATION Update 1.1.x with the suggestions above so the text matches.

RECOMMENDATION Should each function have its own cell? cell 12 and 13

RECOMMENDATION If we are renaming our model titles, then 1.1.6.1 and 1.1.6.1 should match the titles

**2.1 EVALUATION METRICS**

[10 points]

We are using too many metrics in my opinion. I think we should stick to K-fold CV, ROC curve, and Accuracy. Is K-fold CV a metric versus a method of preparing the data for classification? - Choose and explain your evaluation metrics that you will use (i.e., accuracy, precision, recall, F-measure, or any metric we have discussed).

We need more depth for each metric - Why are the measure(s) appropriate for analyzing the results of your modeling?

If we provide more depth for the item above we'll be good for this item - Give a detailed explanation backing up any assertions.

RECOMMENDATION

Both response variables (Crime Solved and Perpetrator Demographic) for both models are categorical. K-fold cross validation (CV) will be used to evaluate the effectiveness of the classification prediction algorithm and measured using ROC Curve, Accuracy, and Sensitivity & Specificity.

RECOMMENDATION Made some changes to these items.

1. \*\*\*K-fold cross validation (CV)\*\*\*

RECOMMENDATION Cross validation allows for the training and testing datasets to be separated. This prevents the bias of the training data's accuracy from skewing the true prediction capability of the model.

2. \*\*\*ROC Curve\*\*\*

RECOMMENDATION Receiver Operating Characteristics curve determines the prediction quality with respect to the predictors. This allows the investigator to reduce dimensionality and complexity, while maintaining a high quality model. A healthy ROC curve, pushes towards the top-left side both for positive and negative classes. The number that tells us the quality of the curve is the Area Under the ROC Curve (ROC AUC) score. We care about both positive and negative classes equally, thus ROC AUC is a good metric.

3. \*\*\*Accuracy\*\*\*

RECOMMENDATION Accuracy will determine the model's predicting capabilities. Accuracy determines the overall % of correctly classified observations, both positive and negative. Our data is balanced, so we won't have an artificially high accuracy score because observations were classified as the majority class.

RECOMMENDATION Drop????

4. \*\*\*Sensitivity and Specificity\*\*\*

While accuracy determines the model's overall predictive capabilities. Predicting classes accurately should also be emphasized. An example would be a skewed dataset which has 80% class A and 20% class B. If the model classifies all records as class A, an accuracy of 80% initially appears promising, until realizing that no records will be correctly predicted as class B. Sensitivity and specificity will ensure the accuracy of the class distribution.

RECOMMENDATION Drop???? We don't have any plots that show the F1 score and the threshold

5. \*\*\*F1 Score\*\*\*

The F1 score is the weighted average of precision and recall. F-measure is a number between 0 and 1 where closer to 1 is better and approaching 0 is worse. It overcomes the limitations of accuracy whenever false positives and false negatives are not about equal or symmetric.

2.2 TEST/TRAIN SPLIT

[10 points]

perfect - Choose the method you will use for dividing your data into training and testing splits (i.e., are you using Stratified 10-fold cross validation? Why?).

perfect - Explain why your chosen method is appropriate or use more than one method as appropriate.

2.3 MODELS

[20 points]

done - Create three different classification/regression models for each task (e.g.,random forest, KNN, and SVM for task one and the same or different algorithms for task two).

done? - Two modeling techniques must be new (but the third could be SVM or logistic regression).

I didn't see this - Adjust parameters as appropriate to increase generalization performance using your chosen metric. You must investigate different parameters of the algorithms!

QUESTION Did we need to do SVM or logistic? Or is it alright that we did three new models?

RECOMMENDATION Naive Bayes is mispelled as Niave bayes in multiple places.

The models used are --

1. Random Forests

2. K-nearest neighbors (KNN)

3. Naive Bayes Random Forests is considered as an ensemble of decision trees.

RECOMMENDATION Random Forests algorithm can be summarized in following steps:

1. Randomly select n samples from the training dataset with replacement called bootstrapping.

2. Make a bunch of decision trees from the bootstrap sample. At each node:

a. Randomly select d features without replacement.

b. Split the node using the feature that supplies the best split according to the objective function, for example by minimizing the gini score.

3. Repeat step 1 and 2.

4. Aggregate the prediction by each tree to assign the class label by majority vote.

RECOMMENDATION KNN applies classification through majority rule and distance calculation. If k is equal to 3, for example, the KNN algorithm will calculate the 3 nearest data points (depending on the distance method) and assign the instance in question to the majority class. Due to the nature of the method, its best to select an odd K, as opposed to an even, to prevent the likelihood of a tie.

RECOMMENDATION Naive Bayes is a probability method which applies the assumption of independence. An example would be deciding the probability of an outcome, given a known prior, in which the highest probability is applied to the instance in question. Due to classification 2 not being binary nor numerical, an ROC curve is not generated for analysis. Two Naive Bayes methods were used -- Guassian and Bernoulli.

RECOMMENDATION Should we create a table for each classification model with the metrics we said we would be using?

<https://analyticsindiamag.com/7-types-classification-algorithms/>

Classification 1

Model AUC ROC Accuracy Sensitivity & Specifity F1 Score

Random Forest .67 .66 didn't see this? .67

KNN .64 .71 didn't see this? .66

NB .68 .72 didn't see this? .65

Same for Classification 2

https://hub.packtpub.com/implementing-3-naive-bayes-classifiers-in-scikit-learn/

RECOMMENDATION We should drop Gaussian since it is only useful when working with continuous values. We should drop Multinomial for model 1 since it is good when there are n elements. It will work well for model 2. Bernoulli is the most optimal for model 1 since it assumes only two values (0 and 1).

RECOMMENDATION 2.3.1 MODEL 1: Random Forest - Classification 1 and Classification 2

RECOMMENDATION I thought that we were not using precision and recall? If we decided to keep F1 score, then we should use it as the measure and not precision and recall.

Reviewing the precision and recall, there is a significant percentage of outcomes which are high, showing that there may not be sufficient data for an accurate model with high sensitivity and specificity. In addition, model tuning may be needed.

Another likely reason for the accuracy, recall, and precision to be low is that there may be classes in the perpetrator demographic concatenation that have low frequencies, and the model is not able to properly classify.

RECOMMENDATION 2.3.2 There is a DISCUSSION text just hanging out. Remove it?

2.3.3

RECOMMENDATION Naive Bayes is the fastest and least computationally heavy of all models, running in seconds. The methods applied below are both Gaussian and Bernoulli. While gaussian is best applied to continuous variables which follow the gaussian curve, the model continues to be applied for comparison. Bernoulli will treat the classes independently as binary classes.

RECOMMENDATION The last confusion matrix before 2.4 ANALYSIS. Is it needed? We don't explain what we didn't provide any analysis of it. Its a great view and something we should write-up.

2.4 ANALYSIS

[10 points]

Missing - Analyze the results using your chosen method of evaluation.

We have visual, just no explanation - Use visualizations of the results to bolster the analysis.

Missing - Explain any visuals and analyze why they are interesting to someone that might use this model.

RECOMMENDATION several empty cells

2.5 ADVANTAGES/DISADVANTAGES OF MODELS

[10 points]

Almost there - Discuss the advantages of each model for each classification task, if any.

Almost there - If there are not advantages, explain why.

Almost there - Is any model better than another?

Missing - Is the difference significant with 95% confidence?

Missing - Use proper statistical comparison methods. You must use statistical comparison techniques—be sure they are appropriate for your chosen method of validation as discussed in unit 7 of the course.

Should we add a table to do the comparison? The advantages and disadvantages are pretty well documented. Dr. Drew provided feedback in minilab about using a table - <https://www.geeksforgeeks.org/ml-classification-vs-regression/>

<https://www.aiproblog.com/index.php/2019/06/20/comparing-classifiers-decision-trees-k-nn-naive-bayes/>

<https://thesai.org/Downloads/Volume4No11/Paper_5-Performance_Comparison_between_Na%C3%AFve_Bayes.pdf>

<https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222>

<https://www.datasciencecentral.com/profiles/blogs/7-important-model-evaluation-error-metrics-everyone-should-know>

2.6 ATTRIBUTE ANALYSIS

[10 points]

Missing explanation - Which attributes from your analysis are most important?

Missing explanation - Use proper methods discussed in class to evaluate the importance of different attributes.

Missing explanation - Discuss the results and hypothesize about why certain attributes are more important than others for a given classification task.

3. DEPLOYMENT

[5 points]

Almost there - How useful is your model for interested parties (i.e., the companies or organizations that might want to use it for prediction)?

Almost there - How would you measure the model's value if it was used by these parties?

Almost there - How would you deploy your model for interested parties?

Almost there - What other data should be collected?

Almost there - How often would the model need to be updated, etc.?

4. EXCEPTIONAL WORK

[10 points]

You have free reign to provide additional analyses.

Missing explanation - One idea: grid search parameters in a parallelized fashion and visualize the performances across attributes.

Missing - Which parameters are most significant for making a good model for each classification algorithm?