The following is a list of topics/questions we may ask on the oral exam.

1. Machine Learning Landscape:
    a. Describe the difference between supervised/unsupervised/semi-supervised learning
    b. Describe some of the challenges of machine learning.
    c. Describe the difference between online and batch learning
    d. Describe the difference between instance based learning and model based learning. What are the advantages and disadvantages of each?
    e. What is model drift?
2. Pandas/Numpy:
    a. Be able to explain the following terms:
        i. Module
        ii. Class
        iii. Instance
        iv. Receiver
        v. Method
        vi. Argument
        vii. Property
        viii. Static Method
    b. What are some of the ways we can load data/create data in Pandas
    c. Describe difference between .loc and .iloc in pandas
    d. What is the advantage of a vectorized approach over an iterative one in Pandas
    e. What is dtype and how is it used?
3. Feature Engineering
    a. Describe feature engineering
    b. Describe the ways you might encode categorical data
    c. Describe the risks of encoding string data as integers.
    d. Compare one-hot-encoding with label encoding. Why would you not want to just encode a column of string data as a column of integers.
    e. Be able to define the following ideas as related to feature engineering:
        i. Cross feature
        ii. Binning
        iii. Cyclic Values
        iv. Outliers and how you might deal with them
    f. Describe the reasons for scaling your data
    g. Which algorithms require scaling? Which do not?
    h. We may show you pictures of data distributions and ask you talk about them. You need to be able to
        i. Describe the normal distribution
        ii. Describe skewness (right vs. left)
        iii. Describe bimodality
    i. Describe the two types of scaling we talked about; normalization/min-max scaling and standardization

    j. Describe the problems of outliers when scaling data
4. Gradient Descent
    a. Describe the loss function for linear regression
    b. Describe how the loss function is used to find the optimal values for a simple linear regression. We may show you a graph of a loss curve and ask you to describe how gradient descent works with respect to that curve.
    c. What is the learning rate for gradient descent?
    d. What is stochastic gradient descent vs. batch vs. mini-batch?
    e. What is the difference between how we approach the best weights in stochastic gradient descent vs batch?
    f. What are the advantages/disavantages of stochastic gradient descent?
    g. Describe “generalization” as it relates to machine learning. What does it mean when a model “generalizes” well?
    h. What is overfitting?
    i. Why do we need a training and testing split?
    j. What is a stratified split?
    k. What is model regularization?
5. Linear/Logistic Regression
    a. What are p-values and how are they used with respect to fitting a linear regression
    b. What is R-squared?
    c. How do we interpret the coefficients/weights of our linear regression in simple linear regression? Multiple linear regression? What might be the problem with interpretation of coefficients in multiple linear regression?
    d. When do we want to use logistic regression?
    e. What is a multiclass vs multilabel classification problem? What types of logistic regression do we use to answer these problems?
    f. Describe the logistic regression cost function and how it penalizes right and wrong predictions.
    g. What function is commonly used in logistic regression to produce a probability for binary classification tasks?
6. PCA/Correlation
    a. What are the different types of correlation?
    b. Describe the process of PCA. What does PCA result in?
    c. What preprocessing do we need for PCA?
    d. Does PCA produce more or less features?
    e. What is the explained variance ratio?
7. Metrics/Bias Variance Trade Off
    a. Describe how to calculate:
        i. Precision
        ii. Recall
        iii. F1 Score
        iv. Accuracy
    b. What are downsides to using accuracy?

    c. We may give you a scenario like fraud/crime and ask you to describe recall and precision with respect to that scenario. For example, say you are building a model to classify someone as a shoplifter or not. What would precision be a measure of? What would recall be a measure of? If you changed the probability threshold for a positive prediction, what would the effect be on precision and recall?
    d. We may show you a confusion matrix and ask you to describe how you would calculate various metrics.
    e. What is the bias-variance tradeoff?
    f. How will you know that your model has high variance and high bias? How do we deal with high variance models? High bias models?
8. SVM/SVC
    a. What is a support vector?
    b. Describe the kernel trick with respect to SVM/SVC
    c. What is a hyperparameter?
    d. How do we find the right values for hyperparameters?
    e. Describe what the hyperparameter C is and what changing its value does with respect to SVM/SVC?
    f. Describe the hyperparameter gamma is and what it accomplishes
    g. What are the benefits of SVC as opposed to logistic regression?
    h. Do SVCs produce probabilities?
9. Regularization Techniques
    a. What is the objective of regularization techniques?
    b. How do ridge/lasso/elastic net regression accomplish regularization with respect to the loss function?
    c. Be able to describe lasso (l1), ridge (l2) and elastic net regularization
    d. How do we use lasso regression to complete feature selection?
    e. What does the alpha value quantify in these regularization techniques?(what will happen if we set the regularization parameter too high value)?
    f. What does the r value quantify in elastic net regression?
    g. What is early stopping and how does it work as a regularization technique?
10. Natural Language Processing (NLP)
    a. What are some of the challenges with text data and machine learning?
    b. Define the following terms
        i. Tokenization
        ii. Stop Words
        iii. Document
        iv. Ngram
        v. Tagging
        vi. Stemming
        vii. Lemmatization
        viii. Corpus/Corpora
    c. What is the bag of words approach?
    d. How does the TF-IDF approach numerically encode text data?

11. Decision Trees/Random Forests/Bagging/Boosting
    a. For a decision tree, how is the Gini impurity calculated?
    b. What is the loss function for a split for a decision tree
    c. We may show you some data and ask you to walk through how the decision tree algorithm works
    d. How does max_depth work as a hyperparameter?
    e. What is a voting classifier?
    f. What is hard voting vs. soft voting? What might the advantage of soft voting be?
    g. Describe bagging?
    h. What is the difference between bagging and pasting?
    i. What is a random forest?
    j. How does feature importance work with random forests?
    k. What are the two types of boosting we discussed and how do they work?
    l. What are the disadvantages of boosting as opposed to random forests with respect to training time?
12. Neural Networks
    a. Describe how a neural network makes a prediction.
    b. What are the different activation functions we discussed?
    c. What are the trade-offs for a neural network as opposed to some of the other techniques we've learned?
    d. Describe the basic idea of the back propagation method.
    e. What is a convolution?
    f. What is pooling?
13. Other Topics
    a. Explain what data leakage is and what problems it causes.