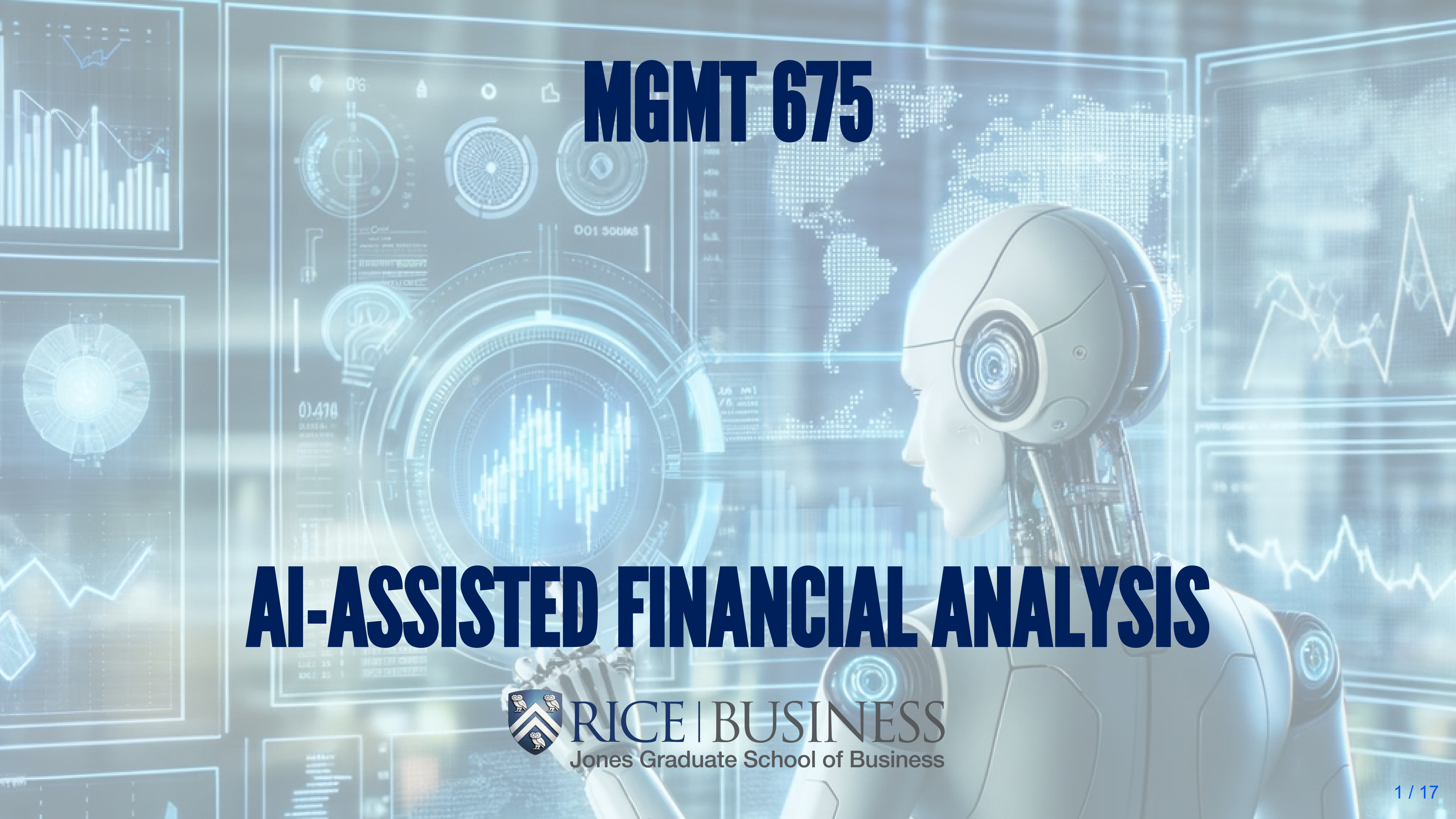

MGMT 675

AI-ASSISTED FINANCIAL ANALYSIS

RICE | BUSINESS
Jones Graduate School of Business

CLASSIFICATION

CATEGORICAL TARGET VARIABLES

- Binary (off/on, yes/no, …) or multiclass
- Forests, neural nets, linear (logistic), plus support vector machines
- Example datasets for today:
  - sklearn breast cancer (binary)
  - sklearn digits (multiclass)
  - loan defaults (binary)

BINARY CLASSIFICATION

- 0 or 1 (negative or positive)
- Predict probability p of positive (e.g., 0.9)
  - Error is p is actually negative (e.g., 0.9)
  - Error is $1 - p$ if actually positive (e.g., 0.1)
- Try to minimize sum of some function (square, log, ...) of errors
- Ultimate prediction is positive if $p > 1/2$.

PREDICTING PROBABILITIES

- For logistic regression (logit), predicted probability is

$$\frac{1}{1 + e^{-\alpha - \beta_1 x_1 - \cdots - \beta_n x_n}}$$

- Higher value for $\alpha + \beta_1 x_1 + \cdots + \beta_n x_n$ produces a higher probability (between 0 and 1).
- Same function is used (usually) for neural networks.

BINARY CLASSIFICATION WITH TREES AND FORESTS

- Predicted probability is class frequency within the leaf.
- Try to split to create pure leaves
- A group is pure if it is all of the same class
- The more mixed a group is, the higher the impurity
- Gini index: like the reverse of diversification
    - Less diversified = less impure = more pure = higher index

EXAMPLE

- Ask Julius to get the sklearn breast cancer dataset and to show the head of the data with the target variable labeled benign or malignant.
- Ask Julius to fit a decision tree and to plot the tree.

ERRORS

- Ask Julius to fit a random forest regressor and to show the confusion matrix on the training data and the test data.

POSSIBLE ADJUSTMENTS

- Can run GridSearchCV for hyperparameters like max_depth
- Can change class weights to emphasize one class (e.g., focus on getting malignant right) or because of imbalanced data (more of one type than another)
- Can change threshold for predicting one class or the other

PREDICTION THRESHOLD

- Ask Julius to plot a histogram of the predicted probabilities for the training data.
- The default rule is to predict the class with the highest probability.
- But you can change the threshold. For example, predict malignant if probability >=25%. Will reduce false negatives (but also increase false positives).

ROC CURVE

- ROC (Receiver Operating Characteristics) curve shows trade-off between different types of errors.
- True positive rate = fraction of positives that are correctly classified
- False positive rate = fraction of negatives that are incorrectly classified
- Ask Julius to show the ROC curve for predicting malignant (i.e., malignant=positive) for the training data.

CLASS WEIGHTS

- Ask Julius to refit the random forest with the class weight of malignant 10 times higher than the class weight of benign.
- Ask Julius to show the ROC curves for predicting malignant for the training data with and without the class weight adjustment.
- From the ROC curves, we can probably see which model we like best and what the threshold should be. Then we can test it.

SCORING FOR GRIDSEARCHCV

- Accuracy (% correctly classified) is usual score.
- But we can also use the true positive rate as the score.
- Or false negative rate or …

ANOTHER EXAMPLE

- Ask Julius to fit a neural network.
- Ask Julius to show the confusion matrix and the ROC curve for the test data.

MULTICLASS

- For trees and forests, predicted probabilities are again class frequencies
- Neural networks produce a separate output for each class. Probability for class i is

$$\frac{e^{\text{output}_i}}{\text{sum over classes of } e^{\text{output}_j}}$$

- Called softmax function.
- Logistic regression is same.

EXAMPLE

- Ask Julius to get the digits dataset from sklearn and to show the head of the data.
- Ask Julius to show the image for the first sample.
- Ask Julius to fit a neural network and to show the confusion matrix on the test data.

DEFAULT DATA

- Start a new chat.
- Download the loan default data from the course website and upload to Julius.
- Ask Julius to show the head of the dataset.
- Build a model to predict defaults!