Slide-1:

- Intro of team
- Overview of presentation
- What is Simplyx
- What is Alteryx

Slide-2:

- Data prep and analytics software
- User friendly, low code, GUI
- Clean and organize datasets

Slide-3:

- Internationally used by blue-chip organizations
- Some brands include…
- Use to improve data prep and utilization efficiency

Slide-4:

- Why Alteryx certification matters
- Improve recognition and understanding of data wrangling
- Alteryx Designer Certification
- Leeds Business school class focuses on Alteryx
- I was a student in that class
- The educational material lacked engagement aspects
- The passing rate was very low
- Peter and Kai collaborated to build Simplyx

Slide-5:

- Dive into Simplyx
- Value proposition
- Unique selling point
- Passing rate increased

Slide-6:

- Simplyx is a series of training exercises that are divided into builts representing each level
- Explanation on the background mechanism of how Simplyx works
- Simplyx features
- Simplyx has data that supports its significance in improving class participation, understanding, and pass/fail rate
- Hand-over to Michael

Slide-7:

- The team:
    - Michael Ghattas (Collaborator)
    - Peter Kinder (Lead)
    - Kai Larsen (DE)
        - Associate Professor of Information Systems in the division of Organizational Leadership and Information Analytics, Leeds School of Business, University of Colorado Boulder.
        - He is a courtesy faculty member in the Department of Information Science of the College of Media, Communication and Information

    - A Research Advisor to Gallup
    - A Fellow of the Institute of Behavioral Science.

Slide-8:

- The domain of the problem relates to an instructional software called Simplyx
- Is used by students to prepare for a certification for a software called Alteryx.
- The goal is to develop a predictive model for student performance on the certification based on Simplyx log data.
- The solution will be used to further support student learning by allowing the course instructor to quickly assess which students are predicted to perform less than desirable on the certification. (Metrics)
- Currently, there are some tools that a professor can use for this, but a single, actionable prediction would be much more effective.
- Additionally, it might be set up that the prediction for individual students is available to reinforce good habits or motivate additional effort.

Slide-9:

- Log data for each interaction a student has with Simplyx.
- This data includes a unique id for each student, problem number, belt number, whether the schema and the problem was correct, and a timestamp.
- The data is utilized to provide the grading metrics and scores essential to the class
- The data were collected with the hopes of being able to improve the product and provide feedback to instructors.
- The data were collected over three semesters from Spring of 2022 through Spring of 2023.
- The owners of Simplyx own the data.
- The data were collected automatically through user interaction with Simplyx.

Slide-10:

- Each belt represents the level/depth of understanding
- One Hot encoding is a method of converting data to prepare it for an algorithm and get a better prediction.
- We convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns, i.e. each integer value is represented as a binary vector.
- Naturall, our data was prepped using Alteryx! (make it fun)

Slide-11:

- We split the data into 80/20 training/testing sets
- We completed EDA to give an overview of the data relationships
- We wanted to start with MLR utilizing a manual and automated stepwise process on the training set
- We identified some features to be possibly selected
- We also utilized AIC/BIC to identify the best fit MLR model we had
- While the results look promising, there was more layers to uncover regarding its generalizability
- MAE: Mean Absolute Error
- RMSE: Root Mean Square Error
- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion

Slide-12:

- With that! As you can see, the results on the test set were imaginary at best
- Can you imagine scoring -100 on a test?! I know we would not want to get a -100 on this presentation if our work is to be assessed by these result

Slide-13:

- Converted percent to rate
- Perform Logit transformation to bound the response and produce results that are consistent with the grading criteria
- Normalization helped reduce bias and visualize the information more clearly as you can see from these plots

Slide-14:

- As you can see now, no student got a score below a 0 (bounded between 0-100)
- Error score still very high
- Preds over estimating rather than under estimate score

Slide-15:

- Model training set performed very well, while the testing sets performed very badly!
- So what's the problem here? Clearly this is indicative of model overfitting
- What can be done?
- Regularization: Address issues of multicollinearity by weighting the features to promote orthogonality
- Dimension Reduction: As many features as samples. Projection, PCA, etc
- Cross Validation: To asses the models generalizability to out-of-sample data
- Ensembles: Further model generalization based on a selection of possible well-fitted models (One-to-many)

Slide-16:

- Ridge Regression: Uses L2-norm regularization (Euclidean distance)
- Alpha: the weight value assigned by the regularization of the features
- Significant improvement to Error scores (300% improv.)
- More balanced distribution between over/under estimating

Slide-17:

- Lasso Regression: Uses L1-norm regularization (Absolute Value)
- Similar to Ridge regression performance with a minor improvement in error score
- We explored multiple methods to address the overfitting issue
- Ensembles methods still needed to improve generalizability of the model
- Tree methods identified as best approach as it addresses non-linear relationships
- Transfer to Peter

Slide-18:

- Boosting: Utilizes weights for training the trees in a similar way as Lasso and Ridge regression methods
- PCA: Dimensionality reduction, orthogonality of features
- GRID: Array based search method to identify the best possible combination of hyperparameters including the number of trees, learning rate, depth, and percentage of features to be included
- Results did not improve from Lasso and Ridge regression based on error scores, as you can see.
- So we considered other methods…

Slide-19:

- Bagging: Utilizes the random selection process that follows bootstrapping methods to reduce variance in the data while optimizing the features selection through a stepwise like process (add/remove)
- As you can see, no improvement to our results based on the error score

Slide-20:

- Random Forest: Instead considering the percent of features for the entire tree, it takes into account the percent of features present in each split when partitioning the the branches of the tree
- Results closer to error scores from Ridge and Lasso regression

Slide-21:

- Finally, we identified two models for a possible best fit, based on the error scores
- As you can see, the results were very close between Lasso and RF
- Can you guess what the tie-breaker is? Over/under estimating

Slide-22:

- Read slides

Slide-23:

- Answer questions