**Group 2 Capstone Presentation: Commuter Transportation Preferences**

*Outline* |*Team Assignments*

**Slide 1-8:** *Introduction and Selecting Data Sources* – Ryan Cook

**Slide 9-12:** *Data Exploration Phase* – Michael Hertel

**Slide 13-20:** *Analysis Phase* – Robert Vandivort

**Slide 21-26:** *Machine Learning Model* – Joey Balaszi

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

*Machine Learning Mode Section* |*Speaker Notes*

**Slide 21 -** *Machine Learning Model:*

“Our team felt the optimal model selection for our dataset was a Linear Regression model. It is easier to use compared to a deep learning model and predicts continuous variables, which is what our target variable is.”

**Slide 22 -** *Machine Learning Model (cont’d):*

“Several benefits of linear regression model is that it’s an easy algorithm to implement, interpret, and efficiently train the model.”

**Slide 23 –** *Initial Machine Learning Model:*

“Once our initial Machine Learning model connected to our SQL database using Python Pandas, we began preprocessing the data to make it more compatible with the model. As Michael mentioned, several columns were dropped and converted to accomplish model compatibility. We also scaled the data and ran the machine learning model over scaled and unscaled data. We discovered scaling our data did not affect our model’s R-squared score. Since our model is linear regression a confusion matrix is not necessary for this analysis.

We then tested the accuracy of our machine learning model by using SciKit-Learn to separate our dataset into a training set with 75% of the records and a testing set of the remaining 25% of records. We trained the model on the training set by giving the model both the values for the input features, such as rain, snow, temperature, as well as the true outcome values for Pedestrian traffic. The model used the input and output values to calculate a weighted coeffecient for each feature. Then the model was given those inputs and had it predict the pedestrian traffic outcome by applying the coefficients to the features. Our initial model had an R-squared score of .6565, meaning future data has a 65.65% chance of fitting the linear regression model.”

**Slide 24 –** *Optimized Machine Learning Model:*

“As Robert mentioned, we experimented with further feature engineering and selection to determine if the accuracy score of our model could be improved.

Ultimately, the model performed best when all the initial features were included, each holiday in the holiday column was left in place, and the day of week was grouped into weekday or weekend. With this feature set, the model had the highest R-squared score at .6585, meaning there’s a 65.85% probability that future data points fitting the linear regression model.”

**Slide 25 –** *What we would change:*

“For future analysis, our team would incorporate more data into our analysis, such as hours. We did have to aggregate the traffic column and condense our date time column in the initial datasets. This resulted in initial starting with 40,000+ rows of data to condensing it down to about 1,000 rows. Again, our team would aim to collect more data for future analysis.”

**Slide 26 –** *Future Analysis:*

“To improve the model going forward beyond the scope of this project, we could obtain more records gathered from the same sources used for this set. We could also look to add more features not available in this dataset, such as data relating to access to public transportation in the area, as that could affect how people choose to commute. Lastly, we would also ask additional questions we hope to answer with the data such as how does gas prices affect commuter habits?”