Final Project

Machine Learning

Segment Deliverables

Segment 1

Team members will be expected to present a provisional machine learning model that stands in for the final machine learning model and accomplishes the following:

* Takes in data from the provisional database
  + A sample database was created, sample\_data.csv, to begin training our machine learning model(s). This data was successfully loaded by sharing the file on GitHub. This allowed the file to read on Jupyter Notebook.
* Outputs label for input data
  + Covid Cases and Housing Cost were assigned as the feature and target, respectively, in order to create a scatterplot. Using the LinearRegression function from the sklearn.learn\_model library, the variables were fit. Next the Covid Cases feature was used to make predictions of Housing Cost. A new scatterplot was created including the Linear Regression Model. The Linear Regression equation is given as , which implies that for each additional Covid Case in a county, the Housing Cost will increase by $0.08.

Segment 2

The team members are expected to submit the code for the machine learning model, as well as the following:

* Description of preliminary data preprocessing
  + Using sample\_data.csv, there are 26 columns. As a precautionary measure, all rows and columns that contain all null values will be removed. The data in this data frame is from year 2017 to 2021. In order to not have Pre-Covid-19 data skew our results, all data prior to the year 2019 will be removed. We believe this will allow us to set a baseline of how housing costs have changed prior to the pandemic.
* Description of preliminary feature engineering and preliminary feature selection, including the decision-making process
  + For the linear regression model, we will use Covid Cases as the feature, and House Cost as the target. In other models, we will remove other columns that we don't feel will be important features in predicting our target.
  + Right now, the target of House Cost is a quantitative variable, which may present a problem with neural networks. If the neural network produces results with low accuracy, we may have to create a new column that represents whether there is a Cost Increase or Cost Decrease from the previous month. This column would be binary and might allow the neural network to have improved accuracy.
* Description of how data was split into training and testing sets
  + When using the basic neural network, the train\_test\_split method from the sklearn library will be used. By default, the train size will be we to 0.25.
* Explanation of model choice, including limitations and benefits
  + Our preliminary choice of model will be linear regression. Using a linear regression model be will beneficial to use if the data continues to stay linear. This will be determined when the entire dataset is imported into the model. If the Covid Cases and House Cost do not have a linear relationship, then another model will be considered. Another limitation of this model is the removal of all features but Covid Cases. These other features that are being left out may play a big role in making predictions of the target using other machine learning models. Other models that are being considered are logistical regression, PCA, random forests, neural networks, and deep learning models.

Segment 3

Students will be expected to submit the working code for their machine learning model, as well as the following:

* Description of data preprocessing
  + Now using the final data set, preprocessing began with removing any rows or columns that contained null values. Using the years 2019, 2020, and 2021 will allow the machine learning models to make conclusions from data pre-covid, covid's peak, and the subsiding of covid. The columns “\_id” will be removed due to it being a string.
* Description of feature engineering and the feature selection, including the decision-making process
  + Due to there being 24 columns in the final data set, not all features should be used. In order to determine which features are good predictors of the target, each feature will be compared with the target and the correlation will be calculated. Any feature that has a correlation coefficient less than 0.20 will be removed before entering the data into the basic neural network.
* Description of how data was split into training and testing sets
  + When using the basic neural network, the train\_test\_split method from the sklearn library will be used. By default, the train size will be we to 0.25.
* Explanation of model choice, including limitations and benefits
  + Using a linear regression model be will beneficial to use if the data continues to stay linear. This will be determined when the entire dataset is imported into the model. If the Covid Cases and House Cost do not have a linear relationship, then another model will be considered. Another limitation of this model if the removal of many possible features. These other potential features may play a big role in making predictions of the target. Other models that are being considered are logistical regression, PCA, neural networks, and deep learning models.
* Explanation of changes in model choice (if changes occurred between the Segment 2 and Segment 3 deliverables)
  + After using the final data in the linear regression model, certain features do show a moderate linear pattern with house cost. Although this machine learning model does give us some insight, additional machine learning models will be used.
  + In order to include multiple features at once, a pivot was made to use a deep learning neural network. Features that showed a correlation of less than 0.2 when compared with house cost were removed before training the model.
* Description of how they have trained the model thus far, and any additional training that will take place
  + Using 18 features, two hidden layers of 8 and 4 neurons with the first layer using the sigmoid activation function and the second layer using the relu activation function, and one output layer with use the sigmoid activation function.
* Description of current accuracy score
  + Having the setting described above, the deep learning model achieved an accuracy of 0.

|  |
| --- |
| * Additionally, the model obviously addresses the question or problem the team is solving. * Using the MinMaxScaler instead of StandardScaler so that values are scaled from 0 to 1, instead of -1 to 1. * Random Forest comparing predictions against y-test. Test correlation and MSE between these two. |

PCA is a statistical technique to speed up machine learning algorithms when the number of input features (or dimensions) is too high. PCA reduces the number of dimensions by transforming a large set of variables into a smaller one that contains most of the information in the original large set.

Segment 4

Students will be expected to submit the working code for their machine learning model, as well as the following:

* Description of data preprocessing
  + Now using the final data set, preprocessing began with removing any rows or columns that contained null values. Using the years 2019, 2020, and 2021 will allow the machine learning models to make conclusions from data pre-covid, covid's peak, and the subsiding of covid. The columns “\_id” will be removed due to it being a string.
* Description of feature engineering and the feature selection, including the decision-making process
  + Due to there being 24 columns in the final data set, not all features should be used. In order to determine which features are good predictors of the target, each feature will be compared with the target and the correlation will be calculated. Any feature that has a correlation coefficient less than 0.20 will be removed before entering the data into the basic neural network.
* Description of how data was split into training and testing sets
  + When using the basic neural network, the train\_test\_split method from the sklearn library will be used. By default, the train size will be we to 0.25.
* Explanation of model choice, including limitations and benefits
  + Using a linear regression model be will beneficial to use if the data continues to stay linear. This will be determined when the entire dataset is imported into the model. If the Covid Cases and House Cost do not have a linear relationship, then another model will be considered. Another limitation of this model if the removal of many possible features. These other potential features may play a big role in making predictions of the target. Other models that are being considered are logistical regression, PCA, neural networks, and deep learning models.
* Explanation of changes in model choice (if changes occurred between the Segment 2 and Segment 3 deliverables)
  + After using the final data in the linear regression model, certain features do show a moderate linear pattern with house cost. Although this machine learning model does give us some insight, additional machine learning models will be used.
  + In order to include multiple features at once, a pivot was made to use a deep learning neural network. Features that showed a correlation of less than 0.2 when compared with house cost were removed before training the model.
* Description of how the model was trained (or retrained, if the team is using an existing model)
  + Using 18 features, two hidden layers of 8 and 4 neurons will use the relu activation function, and one output layer with use the sigmoid activation function. updated scaler and changed first layer from sigmoid to relu. kept output layer as sigmoid. also testing random forests but uncertain of accuracy yet.
  + Changed StandardScaler to MinMaxScaler so values are scaled from 0 to 1 instead of -1 to 1, and changed first layer from sigmoid to relu. The output layer is kept as sigmoid.
* Description and explanation of model's confusion matrix, including final accuracy score
  + Accuracy score improved from 0 to 2.7979e-05.

Additionally, the model obviously addresses the question or problem the team is solving.