Eric Wu, Jacob O'Leary, Caden Summers, Matthew Brady, Yu-Hsin Liao

STAT 380

Professor Yin Tang

4/18/2023

Proposal

Introduction

We decided to use the "House Rent Prediction" (HRP) dataset found on Kaggle, a data science community known for its enriched various data sources. The HRP dataset consists of 4700+ observations and 11 attributes that allow us to make a prediction on the number of rents given a variety of attributes. Here's the list of variables used and explained by the contributor of the dataset:

- **BHK**: Number of Bedrooms, Hall, Kitchen.
- **Rent**: Rent of the Houses/Apartments/Flats.
- **Size**: Size of the Houses/Apartments/Flats in Square Feet.
- **Floor**: Houses/Apartments/Flats situated in which Floor and Total Number of Floors (Example: Ground out of 2, 3 out of 5, etc.)
- **Area Type**: Size of the Houses/Apartments/Flats calculated on either Super Area or Carpet Area or Build Area.
- **Area Locality (Area.Locality)**: Locality of the Houses/Apartments/Flats.
- **City**: City where the Houses/Apartments/Flats are Located.
- **Furnishing Status**: Furnishing Status of the Houses/Apartments/Flats, either it is Furnished or Semi-Furnished or Unfurnished.
- **Tenant Preferred**: Type of Tenant Preferred by the Owner or Agent.
- **Bathroom**: Number of Bathrooms.
- **Point of Contact (Posted.On)**: Whom should you contact for more information regarding the Houses/Apartments/Flats.

Multiple Linear Regression

Starting with the HRP dataset, there are no N/A values, so it is safe to proceed without addressing them. Split the data into Train/Test sets with a reasonably sized split, such as 80/20. Then, we will begin the process to predict the Rent variable based on the other variables in the dataset. First, use the lm() function to create a baseline model with the Training data to predict Rent with every other variable except the "Area.Locality", "Floor", and "Posted.On" variables. "Area.Locality" and "Posted.On" are being excluded because they are categorical variables with too many levels to be relevant, and "Floor" is being excluded because of having too many levels as well as the different levels encoding slightly different information, as the levels for "Floor" are essentially short sentences. Then this model will be used as an input for the step() function to perform variable selection. You can choose whichever direction you wish for this process. After this code is run, make sure to save the final model with the chosen variables. After this, use the predict() function to get the predicted values for Rent in the Test set. Append these values to a copy of the Test set so that the RMSE of the predicted Rent can be calculated. Do this by subtracting the Predicted values from the Rent values, squaring the resulting variable, finding the sum of this new variable, dividing by the

number of rows in the dataset, and finally taking the square root of this value. This will find the RMSE value for Multiple Linear Regression to be compared to the other two methods.

KNN Regression

Once again given that there are no N/A values, we do not have to omit any rows in the dataset. Before starting work on the model, we will have to scale the data (with the scale() function) to ensure that no one variable will have significantly more weight than the others. Out of the numeric variables in the dataset, "Size" is likely to be significantly larger than the others and will heavily affect the distance calculation. After we scale the data set we will separate the response variable from the inputs and we will split the data 80/20 as before. We will use the knn() function from the FNN package and pass the training data, testing data, and response variable to it. To ensure that our model is as accurate as it can be, we will loop through different k-values (1-50) and compare the RMSEs to find the lowest in the group. Once this k-value has been found we will save the ideal model to a variable and compare the RMSE value to the other two models.

Single-Layer Neural Network

Since this is the first time we ever conduct a neural network ourselves, we decided to adopt a similar neural network shown in the book. First, we make sure that the attributes in the HRP dataset are smaller enough so that we can fit a single-layer neural network. Then, we will create the vanilla object with a single hidden layer with 2/3(size of training data + 1) hidden units and a ReLU activation function followed by a dropout layer with rate = 0.4 and an output layer with one unit. In addition, we will add the fitting algorithm with a similar set of parameters but change the metrics to "mean_square_error" since we are measuring the rmse of each model. Afterward, we will fit the data to the model with the same number of epochs and batch size, after making sure that the input data is scaled. We will then collect the final rmse and write up a model comparison using the results from all three models.

Reference

https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset