IEOR 4523 Project Report

Team Member:
Boyuan Xia, Maokang Lin, Joshua Ye
Cining Liu, Jui-Jia Chen
Fall 2023

1. Introduction:

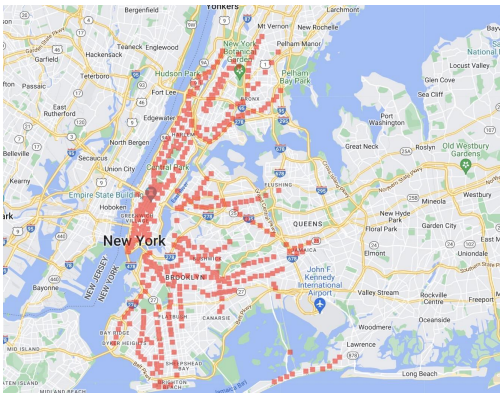
The subway system plays a crucial role in the transportation network of New York City (NYC). As part of our daily commute and exploration of the city, we delved into a comprehensive examination of passenger flows at individual subway stations across NYC. By integrating demographic factors associated with various neighborhoods in the city, we conducted an analysis and prediction of subway traffic patterns. Our goal is to provide insights into the intricate relationship between subway usage and the demographic characteriastics of NYC neighborhoods.

Moreover, considering the impact of the pandemic on subway traffic, leading to a decrease in passenger flow, we extended our analysis to include data from both prior to and following January 2020. This allows us to assess the recovery rate and its correlation with neighborhood factors, providing valuable insights into the complex dynamics of subway usage in the post-pandemic context.

2. Exploratory Data Analysis:

a. And we also calculate the top ten neighborhoods that have the most subway stations, which can be the reference for people who take the subway frequently as an indicator when choosing where to live. We visualize the outcome using Google Map heatmap to know the distribution of the subway stations

Neighborhood	
Financial District	25
Clinton/Chelsea	18
Fort Greene/Brooklyn Heights	16
Bushwick	15
Hunts Point/Longwood	13
Astoria	12
Washington Heights/Inwood	12
East New York/Starrett City	12
Woodside/Sunnyside	12
Rockaway/Broad Channel	11



3. Data preprocessing:

a. Data Cleaning: We obtained two datasets from Kaggle: one focused on subway traffic and the other on the census data of New York City neighborhoods. The subway traffic data spans from February 4th, 2017, to August 13th, 2021, capturing entries and exits at 4-hour intervals across 469 stations. Comprising 4,589,380 entries and 17 columns, this dataset forms the backbone of our analysis. Additionally, we incorporate NYC neighborhood data, associating each station with one of 51 neighborhoods and augmenting our insights with 87 financial and demographic variables.

b. Variables Selection:

i. **Dependent Variable**: We used “Average Flow" as our dependent variable, which represents the average passenger flow at a given subway station, accounting for variations in subway lines and different times throughout the day.

ii. **Data Selecting:** From the 17 columns in the dataset, we narrowed down our focus to 5 key columns: 'Datetime,' 'Stop Name', 'Line,' 'Neighborhood,' and 'Entries' from the NYC subway traffic dataset. Afterward, we sorted the entries values for each stop name and time. Leveraging the 'Entries' column, which signifies passenger flow, we computed the average flow for each subway line at every station during different times.

iii. **Independent variable:** Considering the large number of variables in the NYC neighborhood dataset, we utilized Random Forest to pinpoint variables with an importance index equal to or greater than 0.005.

iv. Furthermore, to investigate potential multicollinearity among the chosen variables (14

in total), we examined their correlation and calculated the Variance Inflation Factor (VIF). To address concerns related to high correlation and VIF, we excluded specific variables: "Students performing at grade level in math, 4th grade," "Moderately rent-burdened households," "Severely rent-burdened households," and "Median household income, homeowners (2018$)." Variables were dropped based on a correlation value surpassing 0.7 and a high VIF exceeding 10.

v. Our final selection comprised 10 columns: 'Housing units,' 'Percent white,' 'Moderately rent-burdened households, moderate income,' 'Pre-foreclosure notice rate (per 1,000 1-4 family and condo properties),' 'Median sales price per unit, condominium,' 'Home purchase loans to LMI borrowers,' 'Refinance loan rate (per 1,000 properties),' 'Single-person households,' 'Residential units within 12 miles of a subway station,' and 'Racial diversity index' as our independent variables.

4. Regression and Modeling

a. Linear Regression:

The first model we used in our analysis was linear regression. We initially applied Multiple Linear Regression using simple Ordinary Least Squares on the data to regress the ten selected features, with a constant added, on average subway flow. Despite obtaining a low R-squared score of 17.6%, all of our feature variables had p-values of zero and were therefore statistically significant. This is indicative of something important missing - that average flow changes every four hours but features do not. In fact, we found thatthere are only six values for the whole year for each station.

Given that, we realized the oversight of not including time as a variable and reanalyzed by splitting the dataset into individual lines and stations. The mean value average flow across four years for each station became the new dependent variable, which should be identical to the mean of the six values of average flow per day. So we want to know if the mean average flow (of 6 values) of each station can be predicted by the station specific independent variables. We look at the data of all 500 stations, using 80% as training data.

b. Decision Tree:

We utilized the decision tree regression model as our second model. Through meticulous data preprocessing, we extracted key features such as the proportion of single-person households, economic status indicators, and demographic and housing characteristics, which are thought to significantly affect subway ridership.

The training of the model took place after splitting the dataset into an 80% training set and a 20% testing set. The decision tree model was chosen for its simplicity and ability to handle nonlinear data. To prevent overfitting, the model depth was capped at three levels. The model's performance(Figure 3) was evaluated using RMSE and R^2 values, with the training set achieving an RMSE of 771.18 and the testing set 937.45, and R^2 values of 0.398 and 0.229, respectively, reflecting the model's shortcomings in data fitting.

The visualization of the decision tree(Figure 4) revealed several insights. For instance, the tree's initial split based on the proportion of single-person households may indicate that living patterns have a direct impact on subway usage. Other economic and housing factors, such as rent burden and property prices, also occupy prominent positions in the tree, suggesting that economic conditions might have predictive value for passenger flow.

However, there are limitations to the model's accuracy. Although the RMSE is below the average flow value, the model's inadequacy in explaining data variance points to a deficiency in capturing linear relationships and temporal dependencies. Notably, the model failed to effectively account for temporal factors, such as the peaks and troughs in flow throughout the day or differences between weekdays and weekends. These shortcomings likely led to the model's diminished performance on the test set, indicating limited predictive ability for unseen data.

c. **Random Forests:** The third model we use is Random Forest, which has higher accuracy than a single Decision Tree. And for this part, it is worthwhile to pay attention to the parameters: number of trees and max depth of each tree. So here we use RandomForestRegressor to run our model, and the range of number of trees is from

100 to 300 and max depth is from 5 to 45. For the number of trees, we find that the model of 100 trees performs best with accuracy > 0.330. It means more trees, lower accuracy, which may be caused by overfitting.(Figure 5) And for max depth, The model with max_depth around 10 performs best with accuracy > 0.360, and there is no obvious difference when max_depth > 10. (Figure 6)This implies there's a non-linear relationship in our data.

d. **Neural Network:** With all kinds of parameters and hidden layers, Neural Network is a complex model which has the best result for our project. In this section, we still focus on the influence of parameters: solvers, activation functions and hidden layers. (Figure 7)

We examined the accuracy of different solvers under different numbers of neurons, and found that both Relu and Tanh activation functions have relevant high accuracy, which means this kind of activation mechanism may be more useful than Logistic. And for different solvers, we found 'lbfgs' works best. (Figure 8 9)

To select an activation function from Relu and Tanh, we also test the learning curve of the functions, which can help identify the fitting condition of our models. Logistic and Tanh have relevant better fitting ability, while Relu is quite underfitting. It may be caused by a gradient vanishing problem in negative value regions.(Figure10 11 12)

In order to discover the impact of both the number of neurons and hidden layers, the examination on neurons is divided into two groups: one hidden layer and two hidden layers. From the result of one hidden layer, we find around 180 neurons the model works best. After adding another hidden layer, the second layer with 50 neurons stands out with an accuracy around 0.38. (Figure 13 14)

e. **Time Series Model:**

The concluding model featured in this report is the time series model, distinguished by its significant deviation from the others. Within the overall dataset, each station's flow contributes to an individual time series. Rather than conducting time series analysis on the entire dataset as a whole, we deemed it more practical to look into the time series of each station independently. Two specific stations, namely the 1st Avenue Station of the Canarsie Line and the Broad Street Station of the Jamaica Line, were chosen for analysis. Additionally, we transformed the data from an hourly to a daily format for enhanced utility.

Attempts to create a time series of Average Flow revealed consistent values daily from 2017 to 2021, making predictions meaningless. Therefore, we chose Normalized flow as our dependent variable for this model, which we found more suitable. The differences between average flow and normalized flow for 1st Avene station are clearly displayed in the diagrams.

Upon reviewing aggregate data, we identified significant differences in normalized subway flow before and after COVID. After all, when the virus hit in March 2020, lockdowns and social distancing kept people from travelling on the subway. We constructed two daily ridership models for each of the mentioned stations, one starting in 2017 and the other in 2020, highlighting the daily seasonality across a week.

We applied seasonal ARIMA models to the four time series we have - 1st Avenue before and after Covid, and Broad Street before and after COVID. In both cases, we used the final 28 days as testing data and everything before that as training data. And the four fitted ARIMA models produced different results, whether it be the number of autoregressive terms, moving average terms, or integrated terms. This shows that ARIMA models vary for individual stations before and after COVID, and also differ among different stations in the same time period.

In essence, we found that the ARIMA models changed quite significantly with COVID for each station.

5. Analysis and Conclusion:

a. Analysis on regression

Through Regression, we find that some factors, including Housing units and Single-person households, are significant and can influence the flow in the subway effectively.

1. Single-person households (positive): Single-person is more likely to take the subway rather than driving, which may be the reason that this factor can improve the subway flow in a positive way.
2. Housing units (positive): The number of houses in that neighborhood. It is reasonable that more residents can bring more subway passengers.
3. Residential units within 12 miles of a subway station (positive): The same as Housing units.
4. Pre-foreclosure notice rate (positive): For people who cannot afford their mortgage, it is most possible that they are facing financial difficulties, and this may encourage them to take the subway and result in a positive impact on the subway flow.
5. Moderately rent-burdened households, moderate income (positive): The rent also becomes a heavy burden for these populations, and this can be explained by financial difficulties as well.

b. Analysis on Models

For the four models we use for modelling, the neural network works best with an accuracy of 0.386. However, the R Square is not satisfactory, the prediction ability of our model still needs more improvement. To solve this problem, it is expected to observe a non-pandemic period after COVID-19 to remove the impact of pandemic. Besides, performing more feature engineering on our data and taking deep learning into models are also promising approaches.

Models	Linear Regression	Random Forest	Neural Network	Decision Tree
R Square	0.343	0.339	0.386	0.229

6. Overall Conclusion and future improvement

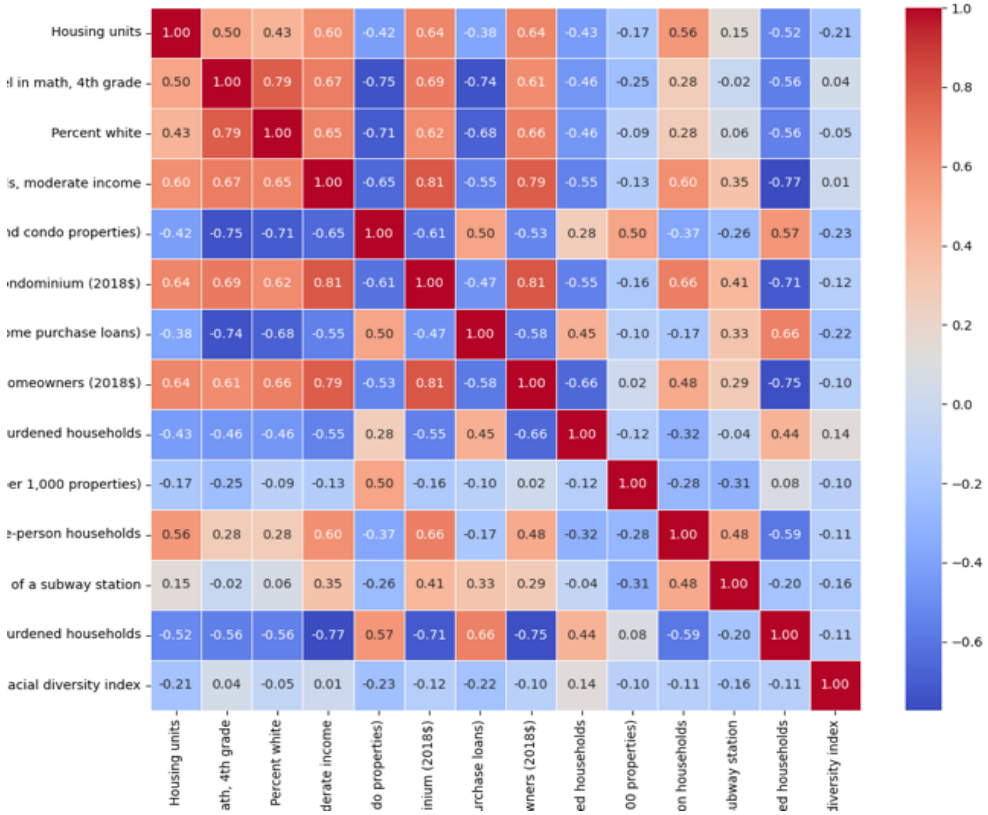
1. Implementing real-time data streams, such as current weather conditions, special events, and traffic updates, could significantly enhance the predictive power of our models. This would allow for more dynamic and responsive forecasting.
2. For the linear regression model, we can consider implementing techniques like Ridge or Lasso regression to prevent overfitting and handle multicollinearity. As for the decision tree model, we can tune different tree depths to balance between underfitting and overfitting. Moreover, in the random forest part, we need to ensure that the trees in the forest are not highly correlated to maintain the diversity in the model and increase the accuracy.
3. Overall Conclusion: Our findings highlight the importance of the subway's proximity to residential areas. Therefore, investing in more accessible subway stations and improving connectivity in underserved neighborhoods could boost usage and convenience. Furthermore, we can provide some suggestions for subway station infrastructure development. For example, expanding capacity in overcrowded stations or optimizing train schedules to meet fluctuating demand. As for the neighborhoods identified as having lower subway usage, the government can launch targeted campaigns to encourage public transportation, which also makes New York City more sustainable.

APPENDIX

Figure1: variables definition

Independent variables	Definition
Housing units	The number of houses in that neighborhood
Percent white	The percentage of whites in that neighborhood
Moderately rent-burdened households, moderate income	For moderate-income households, the percentage of them who pay 30-50% of their income on rent
Pre-foreclosure notice rate (per 1,000 1-4 family and condo properties)	The number of families who failed to make payments on their mortgage for every thousand families
Median sales price per unit, condominium	The median price of a condominium
Home purchase loans to LMI borrowers	For residents who have income no more than 120% of median income, the percentage of them who have home purchase loans
Refinance loan rate (per 1,000 properties)	The number of properties that has a refinance loan per 1000 properties
Single-person households	The percentage of residents who live in a single-person household
Residential units within 12 miles of a subway station	The percentage of residential units within 12 miles of a subway station
Racial diversity index	How likely two people chosen at random will be from different race and ethnicity groups

Figure2: Variables Importance and Correlation Matrix



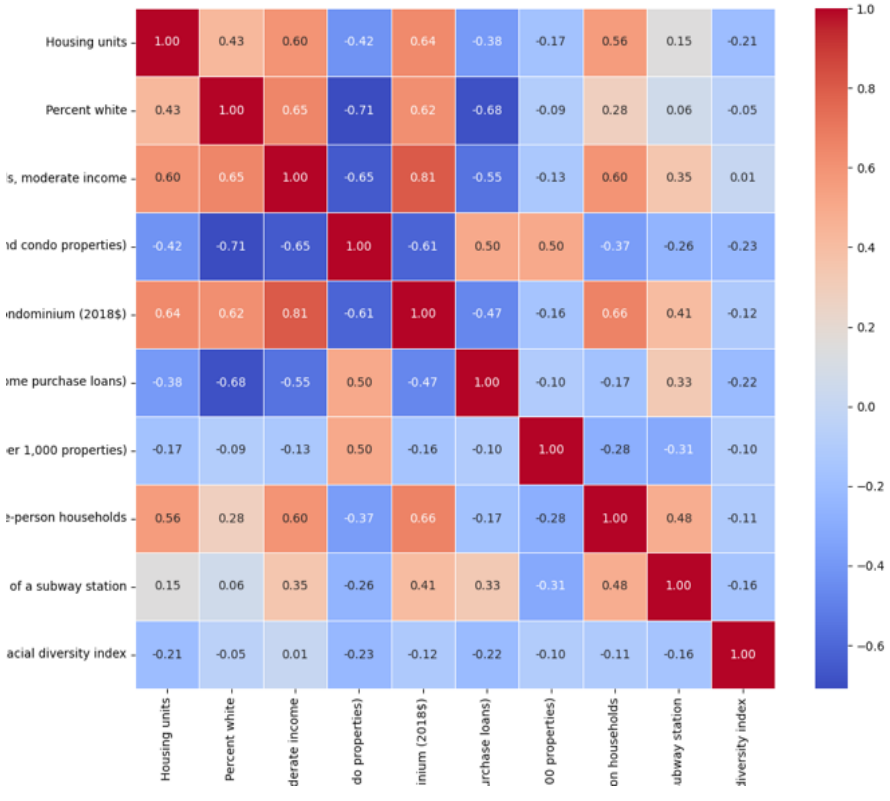


Figure3:

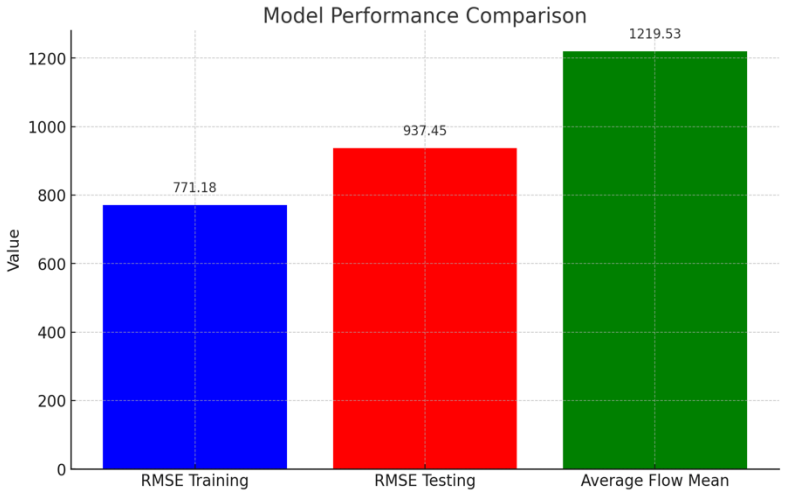


Figure 4:

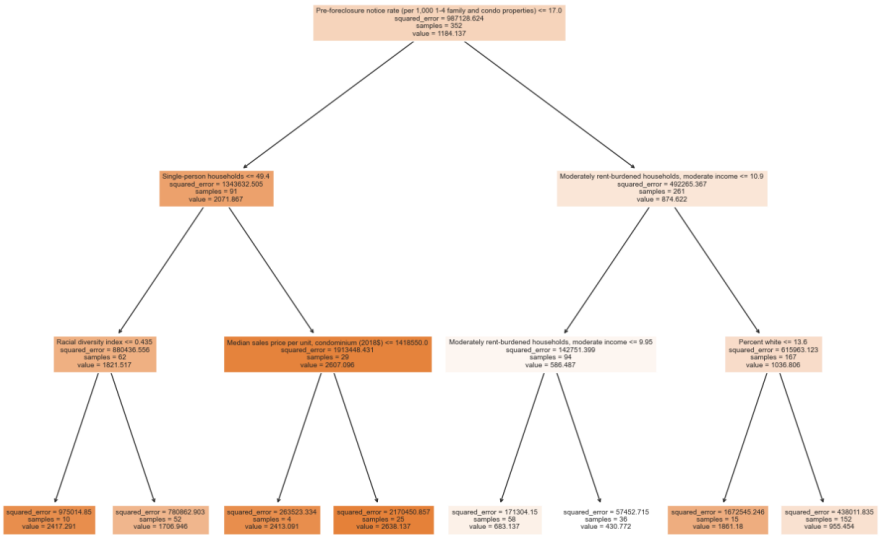


Figure 5:

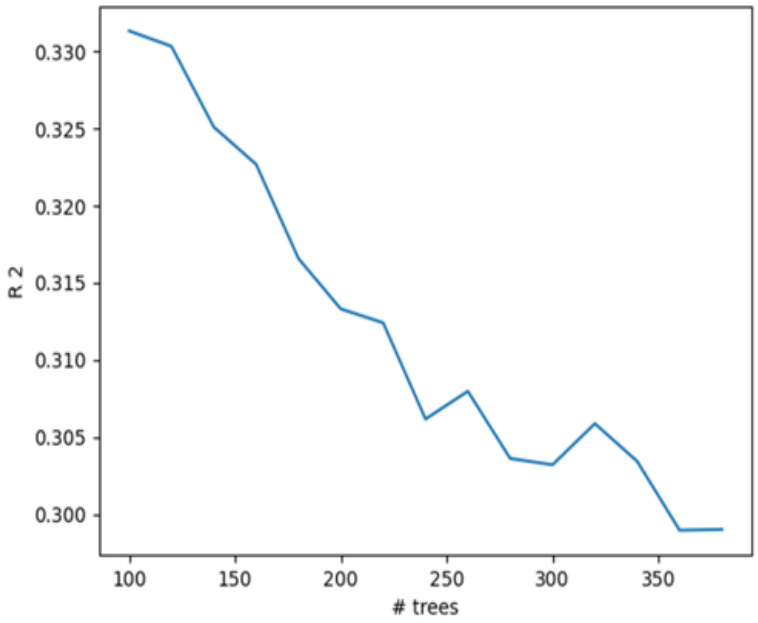


Figure 6:

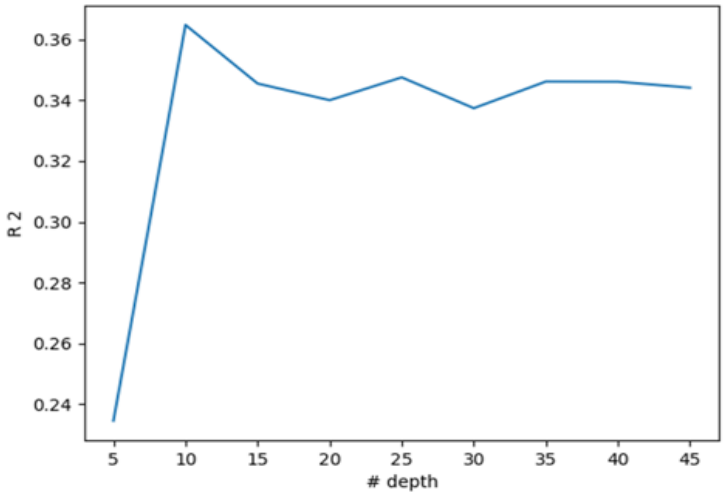


Figure 7:

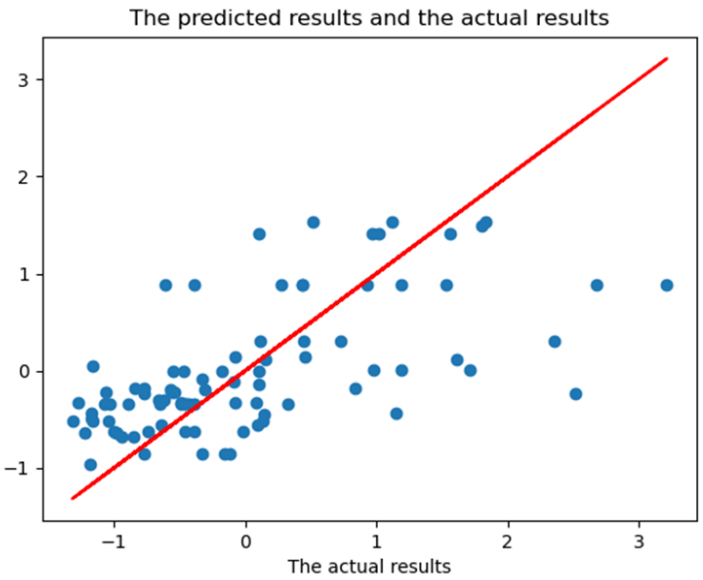


Figure 8:

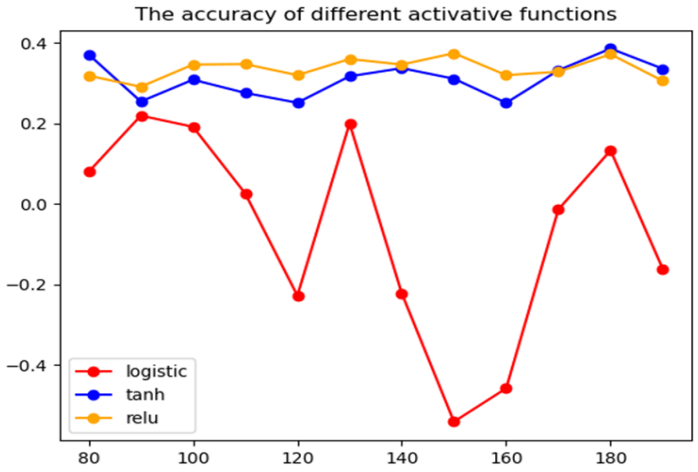


Figure 9:

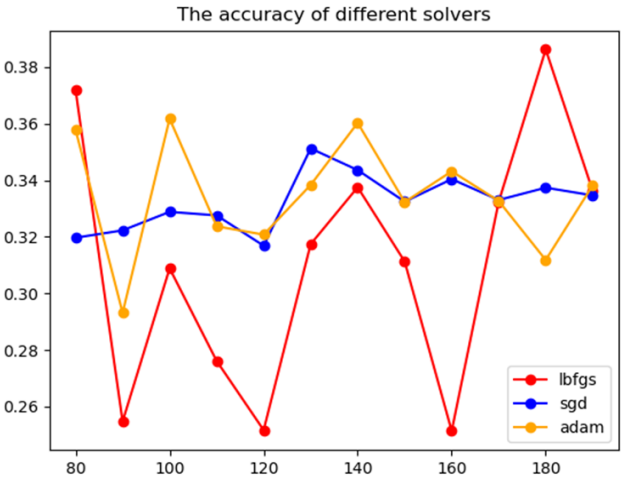
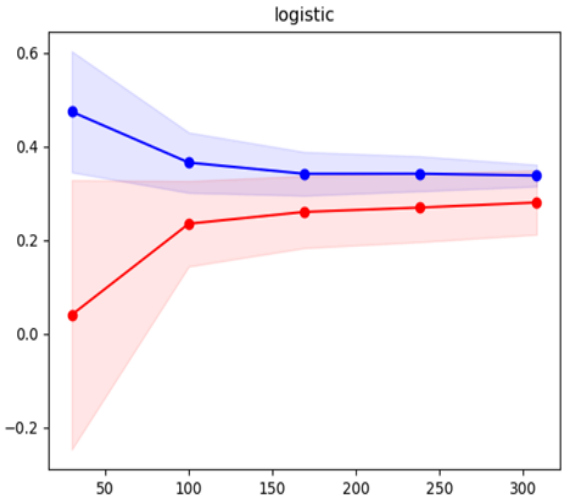
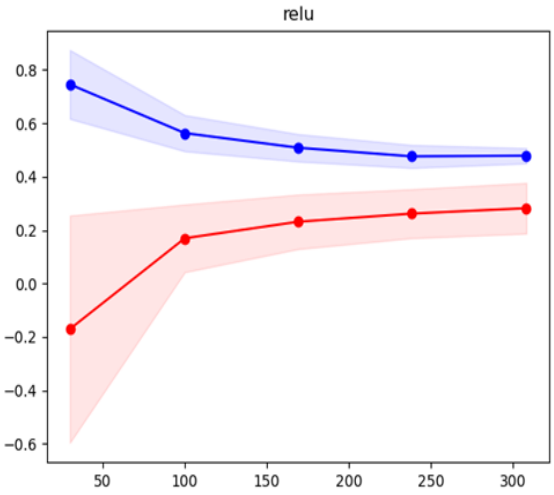


Figure 10 11 12:





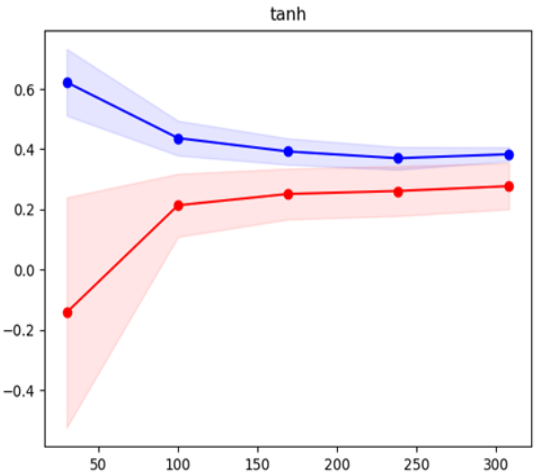


Figure 13:

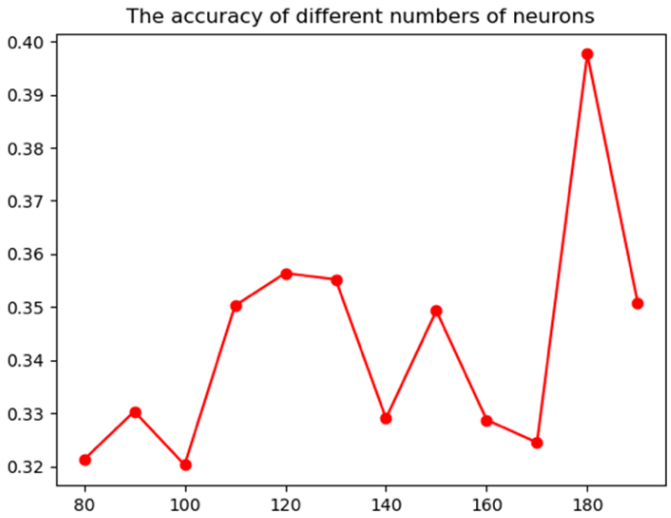


Figure 14: