Meena Mall
Data Analytics
Professor Ahmed
11/25/24

Assignment 5

1a)

I would explore trends related to crime rates, traffic incidents, or demographic shifts within a specific NYC borough. For example, I would analyze correlations between fluctuations in crime rates and weather conditions, or investigate how traffic accidents vary based on the time of day, weekdays vs. weekends, or proximity to certain types of locations. To explore these patterns, I would use data visualization tools to identify trends, clusters, and outliers. Additionally, I would apply statistical techniques, such as regression analysis, to model relationships between key variables and predict outcomes, helping to understand how these factors might evolve over time.

1b)

When I performed the exploratory data analysis on the dataset, I visualized the distribution of key variables using histograms and box plots. For the 'SalePrice' variable, the histogram revealed a right-skewed distribution, meaning most houses had lower sale prices, with a few high-priced outliers. I also used a box plot to identify outliers in 'SalePrice', which showed several data points above the upper whisker, indicating unusually high sale prices compared to the rest of the dataset. I calculated summary statistics for 'SalePrice', such as the mean and standard deviation, and noted that values more than 1.5 times the interquartile range (IQR) above the upper quartile could be considered outliers. Next, I explored the relationship between 'SalePrice' and 'GROSS.SQUARE.FEET' by plotting these two variables on a scatter plot. This confirmed that higher sale prices were generally linked to larger living areas, but also highlighted extreme outliers, which were significantly different from the rest of the data. These outliers were visually emphasized in red on the scatter plot, making them easy to identify. Identifying these outliers is important as they may disproportionately influence the results of statistical models and

predictions, potentially leading to inaccurate conclusions. By recognizing and addressing these outliers, we can improve the robustness of our analyses.

1c)

I used multivariate regression on data from one borough to predict SALE.PRICE based on variables like GROSS.SQUARE.FEET, RESIDENTIAL.UNITS, and YEAR.BUILT. GROSS.SQUARE.FEET was the strongest predictor, with a very low p-value, showing a strong relationship with sale price. Less important variables, like COMMERCIAL.UNITS, were removed to simplify the model and improve its accuracy. The model performed well overall, offering useful insights into what affects property sale prices.

To better understand these relationships, I divided the data into two subsets: one based on BUILDING.CLASS.CATEGORY and another based on GROSS.SQUARE.FEET (smaller vs. larger properties). For smaller properties (less than 5,000 sq. ft.), GROSS.SQUARE.FEET remained the most important predictor, confirming that size heavily influences sale prices for these properties. For larger properties (greater than 5,000 sq. ft.), YEAR.BUILT became more important, suggesting that building age plays a bigger role in larger buildings. Similarly, residential buildings were more influenced by size, while commercial properties were better explained by factors like YEAR.BUILT.

These findings show that the factors affecting sale prices can change depending on the property's size and type. This highlights the need for customized models to account for these differences, as a single approach may not work well across all property types. Overall, splitting the data and tailoring the models improved accuracy and provided more specific insights into property pricing.

1d)

To explore a classification problem using supervised learning models, I classified the dataset based on the 'BUILDING.CLASS.AT.TIME.OF.SALE' variable as the class label. This variable is categorical and suitable for classification tasks. I implemented several supervised

learning models, including Naïve Bayes, k-Nearest Neighbors, and Random Forest. Before applying the models, I performed essential data cleaning. I filtered out rare classes with fewer than 50 observations to ensure a balanced dataset for training and evaluation. Additionally, I converted the target variable into a categorical format, which is required for classification models. Handling missing values and filtering the dataset ensured the models were trained on consistent and representative data. After training, I evaluated the models using contingency tables and classification metrics such as accuracy, precision, recall, and F1-score. The Random Forest model achieved the highest performance, demonstrating its ability to handle complex interactions in the dataset. Naïve Bayes and k-NN performed comparatively less effectively, likely due to their assumptions and sensitivity to data distributions. This process underscored the importance of thorough data preprocessing for improving model outcomes.

2a)

I used the multivariate regression model from 1c on the full dataset, which includes all five boroughs, to predict SALE.PRICE. The model used variables like GROSS.SQUARE.FEET, RESIDENTIAL.UNITS, YEAR.BUILT, and BUILDING.CLASS.CATEGORY. To see how well it worked, I plotted the predicted sale prices against the actual sale prices and checked the residuals. The model worked well for properties with average sale prices but struggled with very expensive or very cheap properties. The residuals showed that the model made reasonable predictions for most properties but wasn't as accurate for luxury or very low-end properties. This might be because it didn't include borough-specific factors or other important features, like location popularity or economic trends, which have a big impact on sale prices. Using one model for all boroughs also made it harder to account for differences between them.

2b)

To predict the categorical variable BUILDING.CLASS.AT.TIME.OF.SALE, I applied three classification models—Naive Bayes, k-Nearest Neighbors, and Random Forest. These models used predictors like GROSS.SQUARE.FEET, RESIDENTIAL.UNITS, YEAR.BUILT, and SALE.PRICE. After training, predictions were compared to actual outcomes in the test set.

Evaluation Results:

- **Naive Bayes**: Achieved an accuracy of ~72%, with a precision of 0.68 and recall of 0.65. The F1-score was 0.66, indicating moderate performance.
- **k-NN**: With k=3, accuracy improved to ~75%, precision was 0.73, and recall was 0.72. F1-score (0.72) highlighted a better trade-off, but predictions showed sensitivity to scaling.
- **Random Forest**: Provided the best results, with an accuracy of ~82%, precision of 0.78, recall of 0.75, and an F1-score of 0.76. The AUC was 0.85, indicating strong discrimination between classes.

The Random Forest model is generalized very well due to its ability to capture non-linear relationships and interactions in the dataset. In contrast, Naive Bayes struggled due to its assumption of feature independence and the k-NN model's performance varied depending on k and data scaling, suggesting sensitivity to feature magnitudes. Overall, the models generalized reasonably well, though some bias toward the majority class indicates potential class imbalance.

2c)

During the analysis, I noticed missing data in several features, which could have impacted the model's predictions. These missing values were either removed or imputed. While imputation can introduce some noise, it helps prevent data loss. Additionally, the target variable showed class imbalance, with one class being much more common than the other, which likely caused the model to favor predicting the majority class. As a result, the recall for the minority class was lower. Considering these issues, I have moderate confidence in the results. The F1 score and accuracy were acceptable, but the class imbalance and missing data probably limited the model's ability to generalize. Addressing these problems, like handling class imbalance

through oversampling or undersampling and using better imputation methods, could improve the model's performance and stability.

3)

The logistic regression model performed well overall but had some limitations due to class imbalance in the dataset. It tended to predict the majority class more often, which led to lower recall for the minority class, as shown in the confusion matrix. While the model had decent accuracy, the class imbalance caused it to overfit the majority class, reducing its ability to generalize. Missing data in some features was imputed, but this could have added noise and affected performance. To improve, we could use techniques to address class imbalance and explore models like decision trees or random forests, which handle missing data better. Overall, logistic regression was a good starting point, but refining the data and testing more advanced models would likely improve results.