Project Report

Shrey Patel

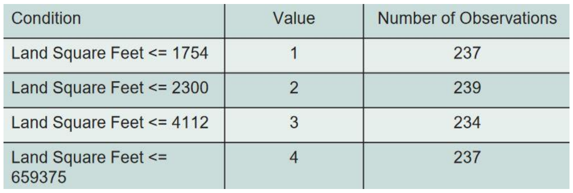
12/15/2021

Introduction

Variables like land square feet area, gross square feet area, number of residential and commercial units, play a huge role in deciding whether a property is worth buying at a given price. Considering the amount of missing data in the sales data set, it is essential to impute or at least semi-impute the missing land square feet area information. In this study, I have predicted the land square feet area for properties using a multinomial Logistic Regression classifier while considering features like the neighborhood, the building class category, block, lot, tax class at sale, building class at sale, and sale price. This study was done to semi-impute (assign a range instead of an exact value) the missing data with respect to the land square feet variable. Post-analyses, I can conclude that the multinomial logistic regression classifier is better at predicting the land square feet area for large buildings and for those that fall under tax class #4.

Data

The data set used in this study is publicly available on the NYC department of finance website. The data contains rolling sales data for a time range between December 2020 and November 2021. There are a total of 20,853 observations and 21 features. For predictive analysis, I created a new target variable, Area from the land square feet variable. The target variable takes four values based on the conditions below:



Condition	Value	Number of Observations
Land Square Feet <= 1754	1	237
Land Square Feet <= 2300	2	239
Land Square Feet <= 4112	3	234
Land Square Feet <= 659375	4	237

The four values represent the inter-quantile ranges. Hence, the labels are well-balanced.

Analyses

Preprocessing Data

Following columns were removed: Borough, Easement, Address & Apartment Number, Gross square feet. Land square feet and sale price variables were converted to numeric. The rest of the variables were converted to factors. The sales date column was converted to a date type column. Following measures were taken to handle missing data: a) all observations with sale price $1000 or lower were removed, b) all observations with missing land square feet area and year built were removed, and c) duplicate rows were removed. The resulting data set consists of 947 observations.

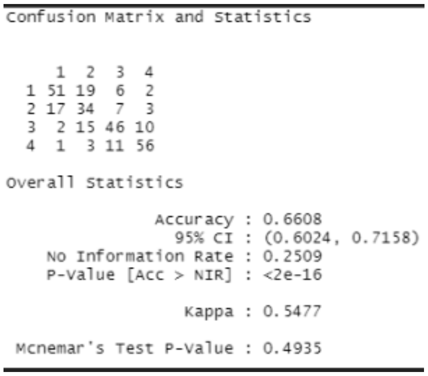
Exploratory Data Analysis

In this study, I have also performed exploratory data analysis to answer the following questions:

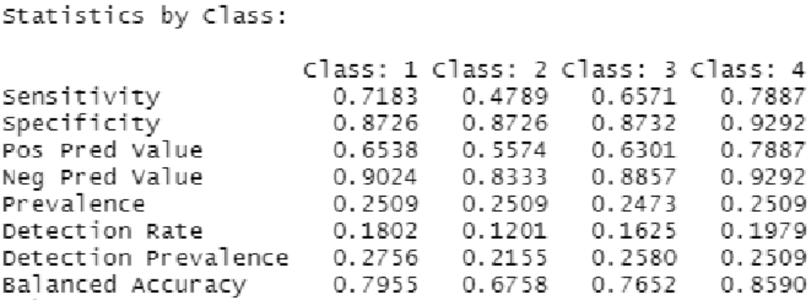
- What is the distribution of sale prices for each tax class?
- What is the distribution of number of sales over time for each tax class? Identify months where the sales peaked or dipped.
- What is the trend with respect to the sale price and the number of sales over time for each tax class?

Predictive Analysis

A multinomial logistic regression classifier was trained and tested using a 70/30:train/test split with 'Area' as the target variable and neighborhood, building class category, block, lot, tax class at sale, and building class at sale as the independent variables. The residual deviance and the AIC of the classifier were 815.033 and 1457.033, respectively. The classifier using maximum likelihood estimation to assign four probabilities (one for each target label) to each observation. The class with the highest probability gets chosen as the predicted class. Below are the results from the analysis.



```
Confusion Matrix and Statistics


    1  2  3  4
 1 51 19  6  2
 2 17 34  7  3
 3  2 15 46 10
 4  1  3 11 56

Overall Statistics

               Accuracy : 0.6608
                 95% CI : (0.6024, 0.7158)
    No Information Rate : 0.2509
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5477

 Mcnemar's Test P-Value : 0.4935
```



```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.7183   0.4789   0.6571   0.7887
Specificity            0.8726   0.8726   0.8732   0.9292
Pos Pred Value         0.6538   0.5574   0.6301   0.7887
Neg Pred Value         0.9024   0.8333   0.8857   0.9292
Prevalence             0.2509   0.2509   0.2473   0.2509
Detection Rate         0.1802   0.1201   0.1625   0.1979
Detection Prevalence   0.2756   0.2155   0.2580   0.2509
Balanced Accuracy      0.7955   0.6758   0.7652   0.8590
```