EXPLORATORY DATA ANALYSIS

August 31, 2023

Amin Zeinali
University of Tehran
Faculty of Mathematics, Statistics, and Computer Sciences

1 Data Overview

The dataset contains 400,988 rows and 62 columns. The columns include phone number, price, status, city, area code, and various engineered features related to the properties of the phone number.

The target variable is ’price’, which represents the price of the phone number. The ’status’ indicates if the number is used or not.

2 Target Variable

In the figure below, the distribution of the target variable is plotted by box plot:

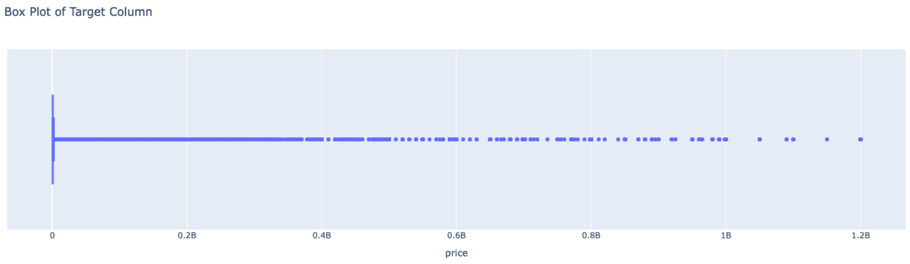

Figure 1: Distribution of the target variable

In order to reduce skewness, the target variable is log-transformed.

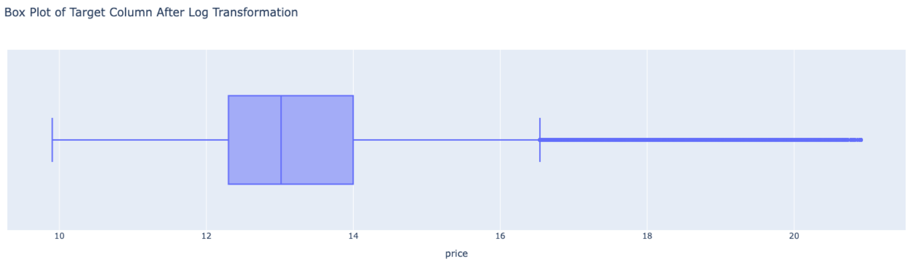

Figure 2: Distribution of the target variable after the log transformation

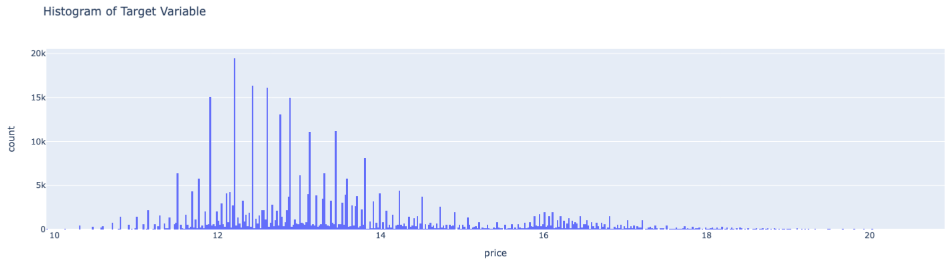

Figure 3: Histogram of the target variable after the log transformation

3 Features

The most common phone number feature is *first_stair*. We can see that *first_stair* and *last_stair* are very close to each other. In addition, there is a significant difference between their count and other features. On the other hand, *all_same* column is 0 for every record so we can drop this column.

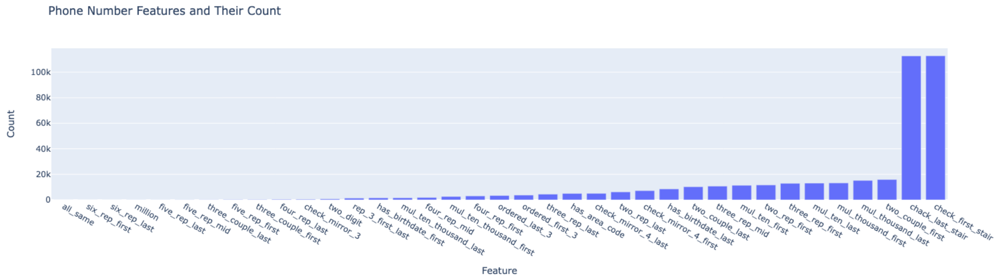

Figure 4: Extracted features and their count

In the following plots, we are going to examine the effect of common and rare features on the target variable using box plots.

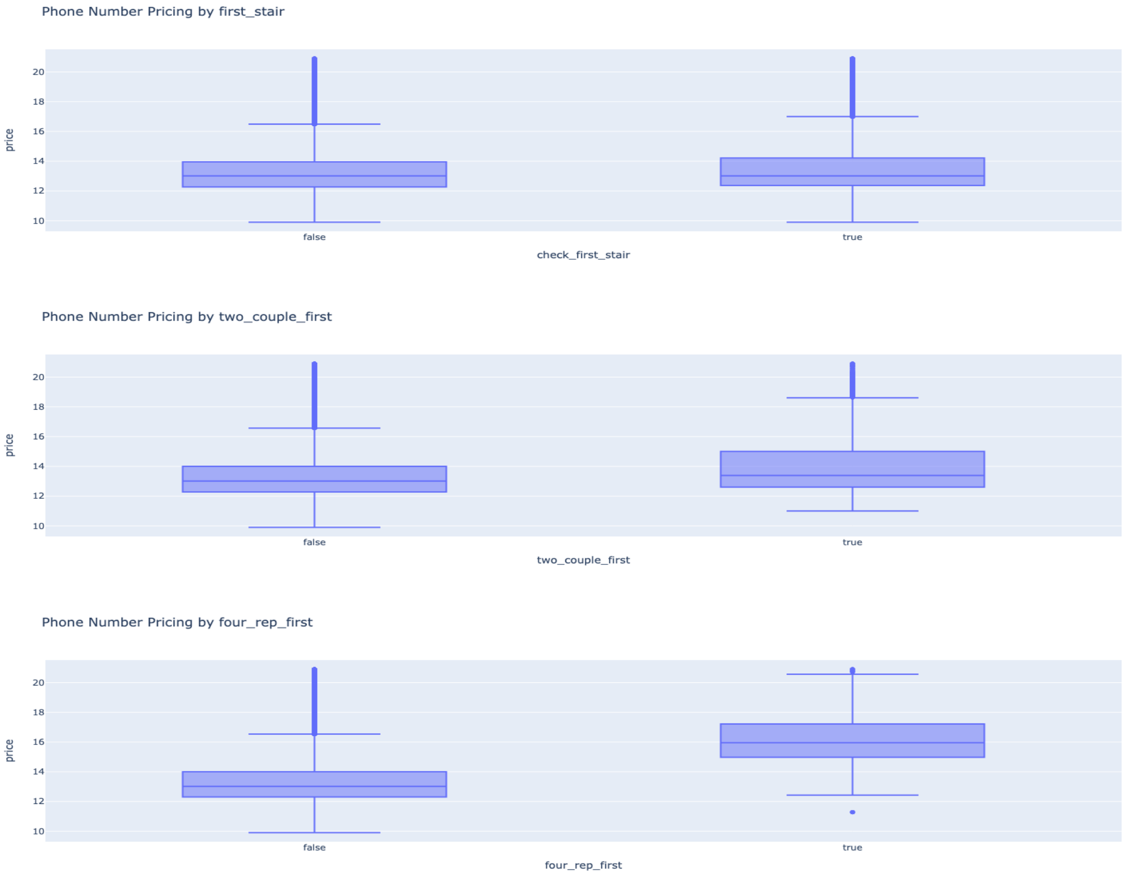

Figure 5: The effect of common and rare features on the target variable.

As you see, rare features like two_couple_first and $four\_rep\_first$ show significant differences in prices.

Now, we analyze the relationship between the first three digits after the area code and prices with a 3D scatter plot.

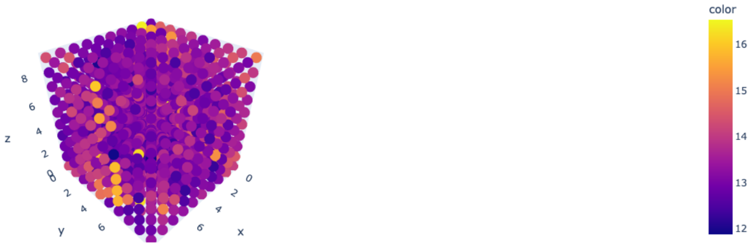

Figure 6: Average price of numbers for different combinations of the first three-digit after area code

The above figure revealed that phone numbers with the first 3 digits as 109, 875, and 871 after the area code are the most expensive on average.

The figure below is the scatter plot of $time$ and the target variable. There is no clear pattern in this diagram, only a slight fluctuation.

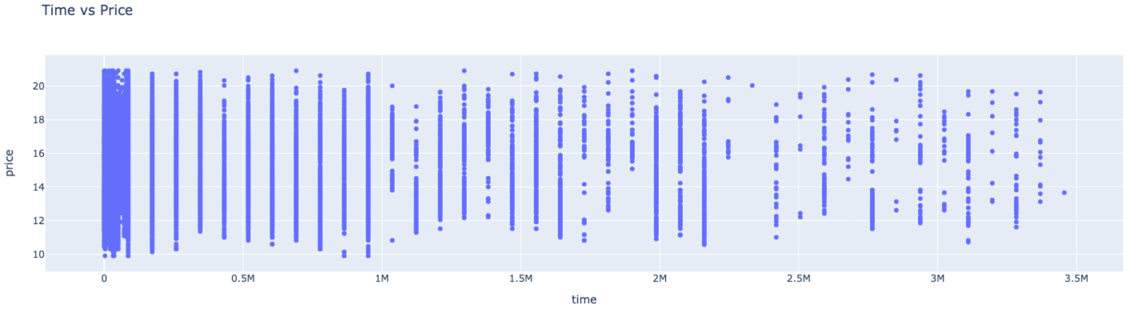

Figure 7: The scatter plot of $time$ and the target variable.

The distribution of the target variable for different $status$ values is illustrated in the figure below.

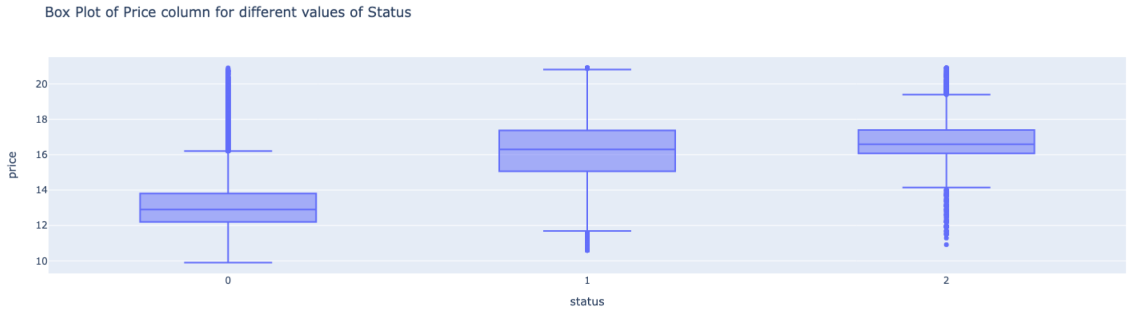

Figure 8: The box plot of the target variable for different values of *status*.

It is clear that used phone numbers tend to be more expensive than unused numbers.

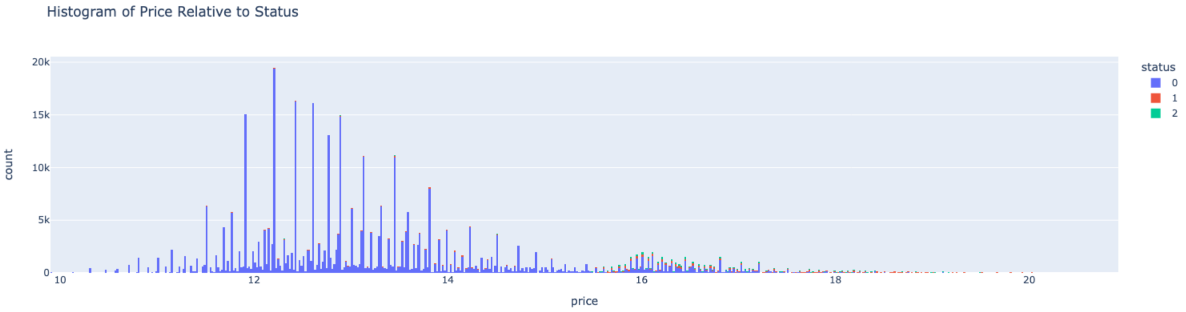

Figure 9: Histogram of *price* relative to *status*.

The bar plot below shows the number of each area code has repeated in the dataset.

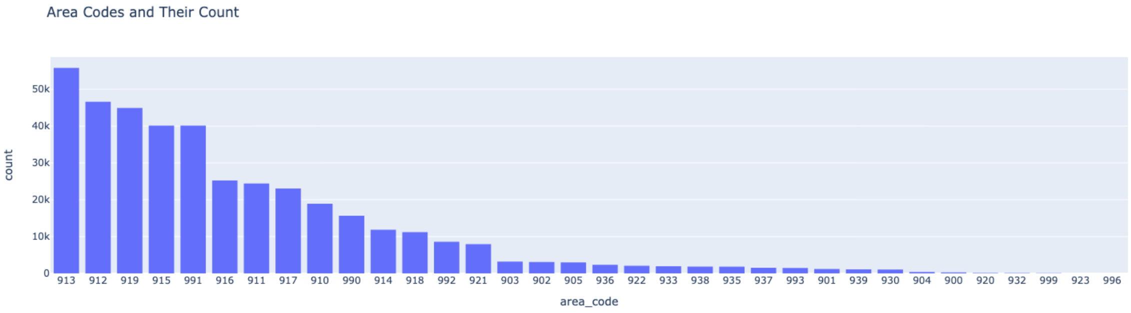

Figure 10: Count of each area code in the dataset.

Over 50,000 numbers have area code 913, while 996 has the lowest count. Now we will investigate its relationship with *price* like other features. Phone numbers with area code 912 have the highest average price, while 993 have the lowest.

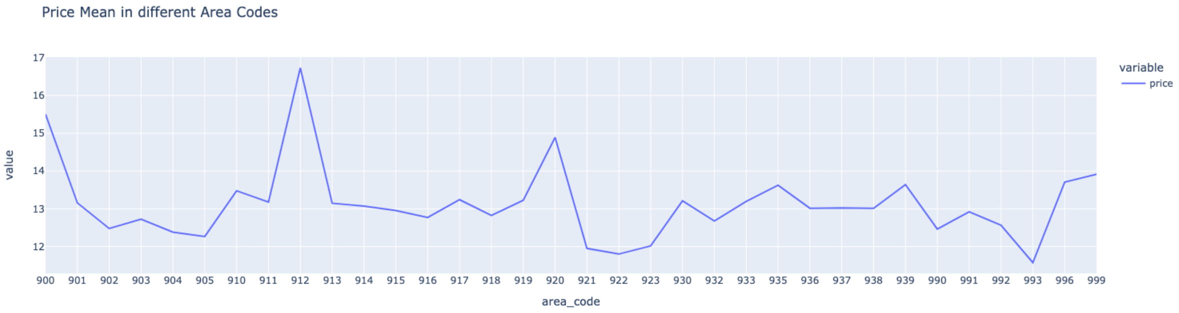

Figure 11: Average price of numbers within different area codes.

4 Summary

In conclusion, the key drivers of phone number price identified during EDA are:

- Usage status
- Area code
- The first 3 digits after the area code
- Rare phone number features

The next steps would be to validate these relationships further through statistical tests and leverage them to build a prediction model for the price. Feature engineering from phone number properties can also help improve model performance.