Seazone Challenge for Junior Data Scientist

For this case, it was used Python, since this is a well established and disseminated language, with easily found documentation and a large number of users. Also, Python is a versatile language, which allows the use of functions to improve code running and memory use.

The project uses a cookiecutter structure, since this is used in most projects of data science. Organization throughout the development of the project was made using git and GitHub.

Data Cleaning

The first step into analyzing the datasets received was importing the data. It was noticed that the types of some of the data were not befitting to the nature of the variable, such as numerical variables listed as string. In order to adjust this information, cleaning functions were made for both datasets, which can be found in the "make_dataset.py" file. In addition, the dataset was limited to the data until March 15th, 2022, that was the date in which the dataset was received.

Features Engineering

Several build features were implemented throughout data analysis and modelling, all of them can be found in "build_features.py" file. Firstly, it was implemented "build_daily_features" in order to determine the number of days between booking and check-in, adding a column to the "df_daily_revenue" dataframe called "reservation_advance_days". The objective was to analyze the pattern of reservation advance.

Here, the "creation_date" was considered as the date of booking, as it differs from the variable "Data Inicial do contrato" in the listings table and it is only present when the listing is occupied ("occupancy" = 1). The advance days number was obtained through the subtraction of "date" from "creation_date" and transforming it into days. Cases in which "creation_date" was greater than "date" were considered to be wrong and turned into "NaN".

For the listings dataset, properties were categorized using numbers, and a column “Quartos” was added to “df_listings”. From the string in the “Categoria” column, the number of rooms was obtained and added to the dataframe in order to model prices and revenues for properties with similar quality, but different number of rooms.

For modelling properties related to time, it was modelled decomposing the date into separate columns, namely “year”, “month” and “day”. In order to do so, timestamp dates listed in the dataframe were converted into datetime format for each feature.

Similarly, it is also interesting to evaluate the effect of the day of the week in the revenue. Thus, a column labeled as “day_of_week’’ was added to the dataframe, using the attribute dayofweek and, by the use of a dictionary, labelling each day accordingly (0 is “Monday” and 6 is “Sunday”). This categorical variable, then, was converted using one hot encoding, therefore preparing it to be used when modelling.

Still regarding to variables related to time, a column was added to daily_revenue dataframe to indicate whether a day is a holiday in Brazil. The function “is_holiday” was implemented using the holiday library. This is useful to help determine the influence of holidays on the revenue obtained in a period.

Another function was made to filter a dataframe containing reservations for New Year’s Eve, in order to determine the advance in which customers make reservations for this date. This function returns the date for a determinate quantile of interest.

Results

1) What is the expected price and revenue for a listing tagged as JUR MASTER2Q in March?

After due cleaning and adjusting of both “listings” and “daily_revenue” datasets, they were merged into a new dataset named “data”. From the “data” dataset, it was observed that there were no listings that corresponded to the “MASTER2Q” category.

Thus, in order to obtain a model which can predict price and revenue based on category and number of rooms, it was necessary to explicit these features in the dataset. The feature “Categoria” was used, from which the number of rooms was obtained by collecting the number contained in the string and the category was then numerically encoded from 1 to 5, being SIM = 1 and MASTER = 5 in agreement with quality order.

Finally, the dataframe was preprocessed in order to fulfill missing data using mean as strategy and also scaled using MinMaxScaler. Then, day_of_week and localization features were treated through one-hot encoding and this dataset was used to train the prediction model.

An instance of the XGBoost regressor was trained using 70% of data, and then the model was applied to test set, using mean average error (MAE) as an evaluation metric. The trained model was applied in a synthetic dataset of listings in category = 5 (MASTER), with 2 rooms for all days of March 2022. The price was estimated through average and revenue was estimated using sum. The **expected price** for a listing tagged as JUR MASTER2Q in March was found to be approximately **R$ 661.14**. In a similar way, the **revenue expected** for this kind of listing is approximately **R$ 4,199.85**.

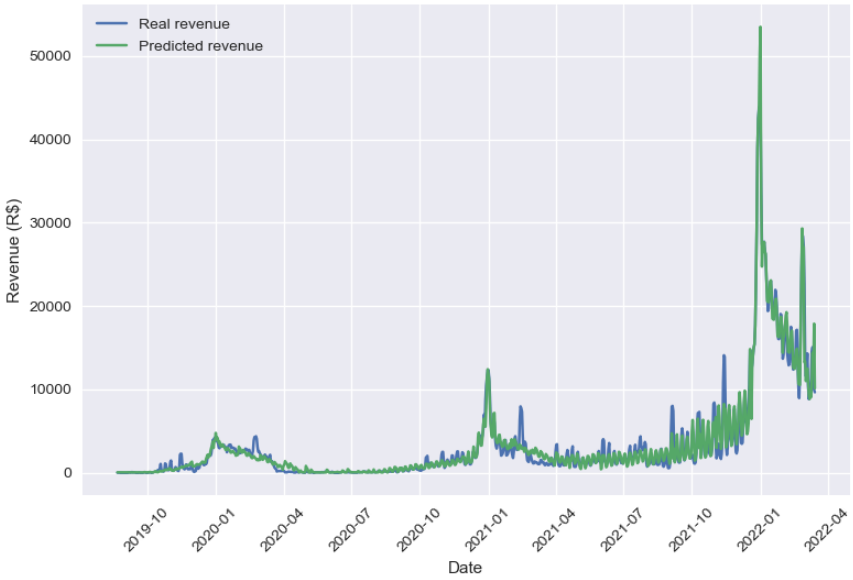
2) What is Seazone’s expected revenue for 2022? Why?

Similarly to question 01, the datasets “listings” and “daily_revenue” were cleaned, adjusted and merged into a dataframe called “data”. To this dataframe it was added a column of “company_revenue”, which is a result of revenue multiplied by commission. The column “date” was decomposed into year, month and day and the day_of_week treated through one hot encoding. The final dataframe contained information on date and company revenue, which are the variables of interest for modelling.

The dataframe was preprocessed in a similar manner, that is, fulfilling missing data with strategy = mean and scaling using MinMaxScaler prior to modelling. In this case, however, it was used MLPRegressor, since the data presents an extremally high level of non-linearity and seasonal peaks throughout the year. The model was also trained using 70% of data and applied to test set

to determine mean average error. It is possible to see in Figure 1 the comparison of predicted values generated by the model and real data.

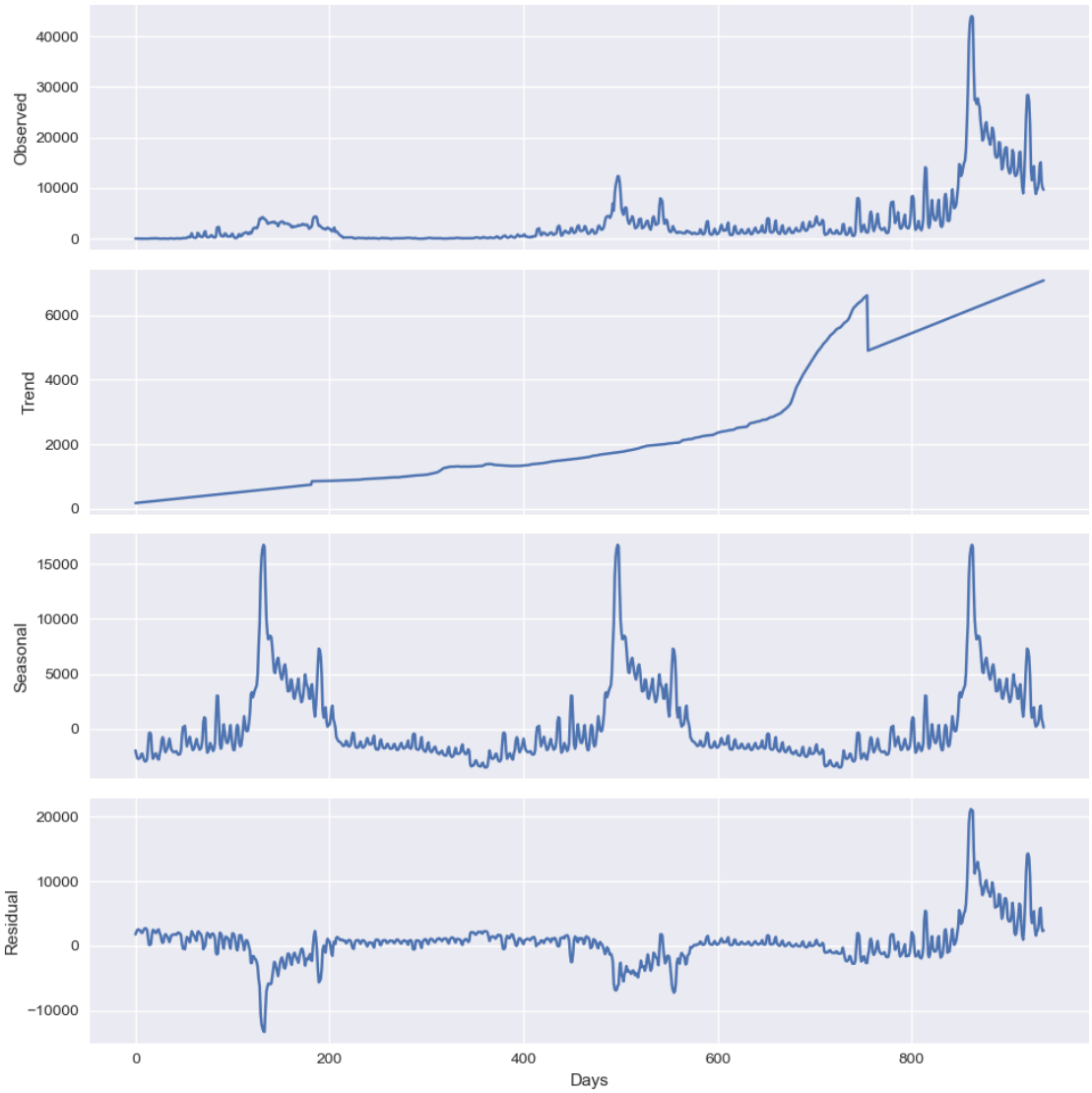
Figure 1 - Predicted and real revenue per date



This model was then applied to dates from January 01st, 2022 to December 31st, and the **expected revenue** obtained was **R$ 4,332,630.25**.

In Figure 2, the company revenue series was decomposed into trend, seasonal and residual parts. The data shows that there is a trend of growth, which began with a sudden rise in revenue around August 2021. Revenue also presents a seasonal behavior, with peaks of interest in the December and January, probably due to Christmas and New Year. A smaller peak is also noticeable right after, around February/March, which is probably due to Carnival. In the same manner as observed in Figure 1, there is a period of low revenue starting in March 2020, which can be interpreted as a direct impact of the measures taken to prevent the spread of Covid-19.

Figure 2 - Seasonal decompose of company revenue



3) How many reservations should we expect to sell per day? Why?

To answer this question, it was used the “daily_revenue” dataset, which was filtered to consider rented listings, that is, with “occupancy” = 1 and “blocked” = 0. Reservations were grouped by day and counted. This dataframe was then treated similarly to the previous dataframes, by applying features to treat date into separate columns through one hot encoding.

4) At what time of the year should we expect to have sold 10% of our new year’s nights? And 50%? And 80?

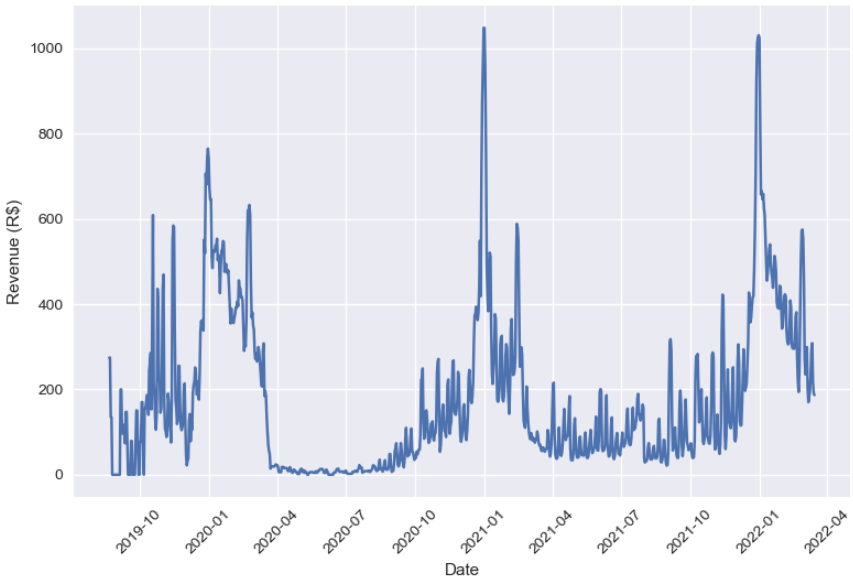
- How can this information be useful for pricing our listings?

…

Complementary Data Analysis

According to the data analyzed, the revenue obtained throughout the year is seasonal, with peaks of interest at the end of each year as can be observed in the graph in Figure 3. It can also be observed that there was a sudden reduction on revenue around March, 2020, which can be explained by the measures taken in order to slow down the spread of Covid-19 pandemic.

Figure 3 - Average revenue per date



Regarding the behavior on reservation advance, it was analyzed the distribution of advance days using the histogram shown in Figure 4. According to this data, approximately 35% of the customers book the listing less than a week before check-in. Similarly, 73% of customers book listings with an advance of up to a month, and only approximately 1.6% of customers do so with advance greater than 6 months.

Figure 4 - Histogram of reservation advance in days