A Work Project, presented as part of the requirements for the Award of a Master's degree in Finance from the Nova School of Business and Economics.

Quantifying Covid-19 impact on Airbnb hosting: Lisbon as a case study

JUSTUS KARL MERZENICH

Work project carried out under the supervision of:
Qiwei Han

04-01-2021

Abstract

While Covid-19 impact on tourism and the sharing economy has proven to be significant by plenty of previous research, data and tools to recursively measure financial impact are missing in the current state of knowledge. This paper aims at quantifying the disease's financial impact on Airbnb prices, bookings and hosting revenues with machine learning. The bottom-up approach used predicts a city's losses at listing level over time and therefore grants leeway to analyzing impact across various dimensions. The city of Lisbon is used to showcase the model's performance and versatility of results.

Keywords:
Covid-19, *Airbnb, Sharing Economy, Data Science, Machine Learning*

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Table of Contents

2. Literature Review

2.1. Impact of Covid-19 on tourism and the sharing economy

The global outbreak (WHO 2020) of respiratory disease Covid-19 (C19) has catalyzed dedicated scientific research like no other event in recent history (Callaway 2020).

Flattening infection curves by curbing mobility and social activities has proved to be crucial in keeping global health care infrastructure intact (WHO 2020), but as research shows, it also comes at great cost to the tourism sector. Yang et al. have measured that throughout prior health disasters, labor and productivity declines have caused significant consumption losses in the industry (Yang, et al. 2020). Since its products and services rely heavily on consumer mobility and social activity (Rocca 2015), the sector is currently in a financial deadlock that is unlikely to loosen up before the end of the pandemic. Planned travel behavior has drastically declined, and planned trips have decreased in duration (Junxiong Li 2020). Karabulut et al. measure negative impact of international pandemics on tourist arrivals in lower-income economies (Karabulut, et al. 2020). The hotel industry has experienced substantial valuation losses (Abhinav Sharma 2020) and the sharing economy in particular has been identified as one of the industries most subdued by C19 (Hossain 2020). Airbnb, one of the world's largest corporate players in tourism and the sharing economy, had to lay off 1,900 employees (i.e., roughly one quarter of its workforce) in early May (Airbnb 2020). Hu et al. state that Airbnb booking activities have declined by roughly 60% and that every doubling in infections implies 4% in booking decreases (Hu und Lee 2020). According to Chen et al., who analyzed short-term (January to March 2020) C19 impact in Sydney, Airbnb hosts' income losses were eight times as large as the company's own losses (Chen, et al. 2020). Evidence for local residents' significant willingness to pay for risk-reduction in tourism has been supplied by Qiu et al. (Qiu, et al. 2020).

2.2. Lack of instruments to quantify financial losses in short-term rentals

While the severe impact of C19 on tourism and the sharing economy is deducible from each aforementioned piece of C19 research, practical and financial implications on hosts, business and economics remain limited. Due to the ongoing nature of the pandemic, some of 2020's earlier research findings might already be obsolete. Almost all above research emphasizes that quantifiable information is either limited, unavailable or outdated. Therefore, decision makers who need up-to-date intelligence on the financial impact of C19 are left in the dark. This holds for governments trying to allocate bailout resources, managers trying to measure their portfolios' losses, as well as individual hosts benchmarking their prices. According to Kock et al., previous research has also missed to emphasize long-term structural and paradigmatic changes induced by C19 (Kock, et al. 2020). This development indicates that less weight should be allocated to point estimates and highlights that models for recursively quantifiable results are required.

A practical example of research in need of bottom-up quantifiable results is Dolnicar and Zare's conceptual paper on regulation. They argue that C19 might be taking over the regulation of excessive Airbnb-caused gentrification because return-driven investors are losing out on idle rooms and shift from short-term into long-term renting (Dolnicar und Zare 2020). Long-term renting is preferred by local city governments because it does not drive rent prices up as much as short-term renting, which has been proven specifically for the case of Airbnb by research from Barron, Kung and Proserpio (Harvard Business Review 2019). According to Dolnicar and Zare, this shift implies that losses from delisted items (those that have been made unavailable to short-term renting during C19) could potentially be viewed as progressive socio-economic development. While their aforementioned paper is crowded with reasonable assumptions, it does not quantify the extent to which losses actually come from cancellations and booking aversion (both of which are harmful to local disposable income) and how much comes from

delisted items (which are assumably absorbed by long-term renting). A model that is able to map losses to these three forces is required.

2.3. Research objective

In order to build upon previous research and shed light on a city's financial C19 impact on short-term rentals, the research goal of this paper is to develop a bottom-up model that takes individual listings into account and is therefore able to recursively map C19 impact to individual listings and their features. Ideally, the model will be able to answer the following research questions deducible from the above literature review given any city:

- **Q1.** C19 impact on supply & demand dynamics since the first lockdown: How many bookings are missing and how are the pricing benchmarks affected?
- **Q2.** What is the magnitude of hosting revenue losses in the city over time?
- **Q3.** How can this loss be deconstructed? How much loss is caused by cancellations, delisted items and aversion?
- **Q4.** What differences in features distinguish a listing's financial impact?
- **Q5.** What areas should city governments allocate most financial aid to? How are the hosting losses geographically distributed?

2.4. Lisbon: Applying the research objective to a case study

In order to demonstrate the model's development and results, the city of Lisbon is used as a case study. Portugal's capital with roughly half a million inhabitants (ANMP 2020) has one of Europe's highest rates of short-term rentals per thousand inhabitants (AirDNA 2020). According to an article published by *Spiegel* (Ginzel 2019), district mayor Miguel Coelho indicated that between 2013 and 2019, the number of short-term rentals offered radically increased from 63 to 5,000 in the area of Alfama, Castelo and Mouraria. According to *Financial Times*, Alfama's local population has decreased from 20,000 to only 1,000 over the past 40 years

(Wisniewska 2019). These statistics imply that short-term rental hosting revenues have become an increasingly important source of disposable income in Lisbon. Since the city is also in an ongoing gentrification clash with investors who drive prices up and thus locals away from the central areas (Warren 2020), diving into the structural changes identified in the literature review (e.g., the share of loss from delisted items that flows into long-term renting) is specifically relevant to the city of Lisbon. Therefore, choosing the city as a case study reveals large scale impact and offers immediate results where they are lacked. While there are other players in the market for short-term rentals (e.g., HomeToGo), Airbnb is by far the most relevant - the company has a 75% market share in Lisbon (AirDNA 2020). The models developed in this paper are therefore exclusively fed with data from Airbnb.

In order to understand and interpret the case study results yielded by the models developed in this paper, broad knowledge about Lisbon's C19 timeline is required. According to Reuters, Portugal recorded its first case of the disease on March 2, 2020 (Reuters 2020). By March 10, merely a week later, partially state-owned airline TAP had already cancelled over 3,500 flights, travel from Italy was suspended. The source also mentions that Portuguese hotel association AHETA had registered a 60% percent cancellation rate by that day (Reuters 2020). In line with research question Q3, identifying actual losses induced by cancellations will shed light on how early money has been lost from bookings being cancelled. By March 12, all schools and nightclubs were closed and on March 15, tourism between Spain and Portugal got officially suspended (Reuters 2020). On March 18, the country officially declared the state of emergency (Portuguese American Journal 2020) and locked the country down entirely. While in April daily infections counted almost a thousand per day, they have plummeted over the summer. In August, daily infections counted only 100 cases in the entire country. Until November, cases have surged again in a second wave and currently count several thousand new cases per day (John Hopkins University 2020). The date of the final analysis is December 19, 2020.

3. Methodology

3.1. Developing a theoretical bottom-up model for revenue loss

As emphasized in the research objective in section 2.3, answering the research questions and identifying relevant dimensions require a bottom-up modeling approach that calculates impact from individual Airbnb listings and nights. While the city itself does not require a variable in this kind of analysis, its sets of Airbnb listings (*L*) and nights (*N)* do. The quintessence in answering Q1-Q5 with regards both to Lisbon as a case study and to developing a generalizable model lies in computing any city's monetary loss per night over time. This loss is denoted as the given city's nightly discrepancy ($D_{L,n}$) between the revenues observed under C19 ($R\_covid_{L,n}$) and the hypothetical revenues as they would have occurred in absence of the disease ($R\_hyp_{L,n}$). Since hosting revenues can simply be modeled as the product of a listings price per night (*P*) and its binary booking status (*B*), the hosting revenue discrepancy $D_{L,n}$ can be defined as in Eq.I (*Figure 1*).

Figure 1*:* Equations to compute estimates hosting revenue loss

Equation I: $$D_{L,n} = \sum_{l=1}^{L} \left((P_{covid_{l,n}} * B_{covid_{l,n}}) - (P_{hyp_{l,n}} * B_{hyp_{l,n}}) \right)$$

Equation II: $$\hat{P}_{hyp_{l,n}} = \theta_1[X1_{l,n}], \quad \hat{B}_{hyp_{l,n}} = \theta_2[X2_{l,n}]$$

Equation III: $$\widehat{D}_{L,n} = \sum_{l=1}^{L} \left((P_{covid_{l,n}} * B_{covid_{l,n}}) - (\theta_1[X1_{l,n}] * \theta_2[X2_{l,n}]) \right)$$

Source: Author

Due to the hypothetical nature of all variables with *_hyp* suffix, their state is predicted rather than observed. Again, observed values are those measured in the real-life C19 scenario, hypothetical values are those that would have been likely to be overserved in 2020 if C19 had not spread. Using the estimators defined in Eq.II, the estimation for the hosting revenue losses can be reformulated as in Eq.III. Note how revenues are not predicted directly but calculated from

estimates for prices and bookings. This approach is chosen deliberately as it is advantageous over a direct revenue model: The latter requires an additional degree of freedom for developing a revenue formula itself and is inaccessible for any exclusive analysis on prices and bookings.

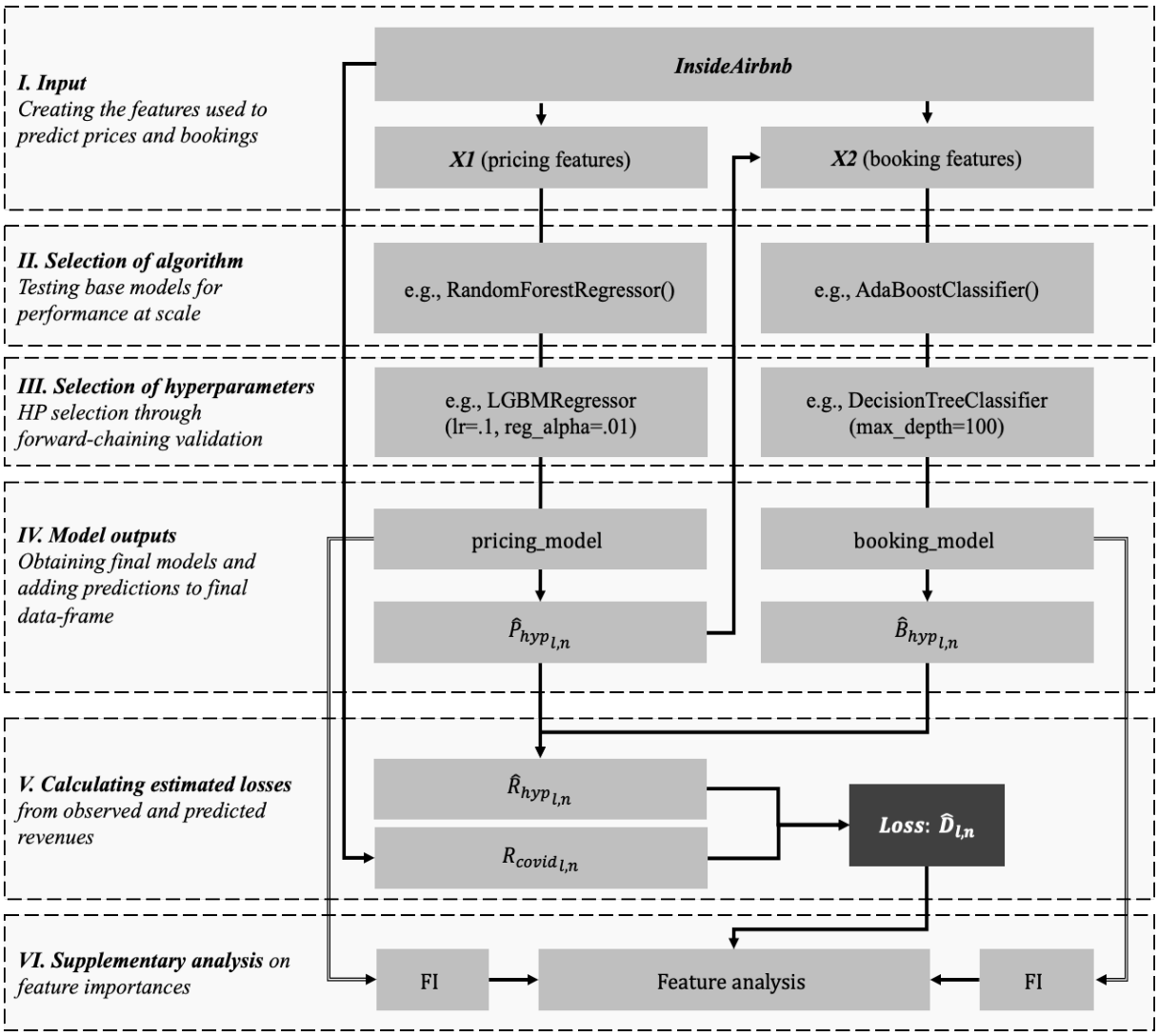
As per Eq.II, a listing's hypothetical price is estimated using θ_1 (pricing_model) given feature set *X1*, its hypothetical bookings using θ_2 (booking_model) given *X2*. While the former takes a qualitatively selected subset (based on value counts, distributions, missing values etc.) of the applied dataset's features, *X2* is simply the combination of *X1* and $\hat{P}_{hyp_{l,n}}$, which implies that booking predictions depend on predicted prices. The fact that bookings might reversely influence price remains neglected. Inherent with this setup, the focus of applying the model to real data will lie with first developing a model to detect a listings hypothetical price (for all nights and listings). Afterwards, the booking model can be developed using previous features together with predicted prices.

3.2. Applying the model for hosting losses to real Airbnb data

In order to apply the theoretical model for hosting revenue losses to real data, the practical framework from *Figure 2* is utilized. In step I, data from publicly available dataset *Inside Airbnb* (Cox 2020) is collected, curated and fed into the pipeline for modeling prices and booking states. The dataset's relevant share for the case study includes over 35,000 listings (*L*) across 678 nights (*N*), is updated monthly and contains anonymized details on listings' properties, locations, reviews and calendars for one year in advance. The data is iteratively scraped from *Inside Airbnb* using a modified version of a web-crawler code supplied by the supervisor (just like the crawler, the entire analysis is written in Python3). An overview of the directory structure used can be found in Appendix 1.1, all packages deployed are listed in Appendix 10. Once the web-crawler has downloaded the files, they are stored in a directory named after the scraping date (the date on which *Inside Airbnb* made the files available). Due to the data's redundancy inherent with the monthly updates and one-year scope, only the latest available

information per night is stored in the final data. While reading the *.csv* files into memory, computational speed can therefore be increased by storing data that looks no further than 100 days into the future. As long as the data is continuously updated by *Inside Airbnb* each month, all data that looks further than 30 days ahead is updated within the next scrape and can therefore be neglected at the current scraping index. Choosing 100 days (i.e., three months) is a safety net for missing months, which do occur across the data.

***Figure 2**:* A practical framework for the development of a bottom-up loss model



Source: Author

The exact subset of variables extracted from *Inside Airbnb* and used for this paper's analyses is displayed in *Figure 3*. A more elaborate data dictionary containing all extracted variables, their data types, shares of missing values, length of value counts and most common values can be found in Appendix 1.5. For simplicity, it is assumed that the presence of C19 does not have an impact on the variables used as features. While for some this assumption does not necessarily hold (e.g., review scores from quarantined guests forced to prolong their stay), most variables should be robust to this context (e.g., latitude, number of bathrooms etc.).

***Figure 3**:* Names of all features used in applying the hosting loss model

calendar.csv.gz	listing_id, date, available, price, minimum_nights, maximum_nights
listings.csv.gz	id, host_response_time, host_response_rate, host_acceptance_rate, host_is_superhost, latitude, longitude, property_type, room_type, accommodates, bathrooms, bedrooms, beds, bed_type, security_deposit, cleaning_fee, guests_included, extra_people, number_of_reviews, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, instant_bookable, is_business_travel_ready, cancellation_policy, require_guest_profile_picture, require_guest_phone_verification, neighbourhood_cleansed

Source: Inside Airbnb

In order to feed the data into predictive models, plenty of data curation, transformation, encoding and imputation is required. Many columns include object-type percentages or currencies read in as strings (e.g., "45,1$" or "50 %"). They are transformed into float-type columns so that all model types can handle them. Columns with binary or ordinary categorical variables are label-encoded in order to limit the number of features created. Categorical variables with no order and longer value counts are recoded using a simple and commonly used one-hot-encoder (Brownlee 2020). The length of excessive value counts is decreased for some columns by automatically mapping contents to contextually similar clusters (e.g., "entire townhouse" and "entire condominium" in the feature *property_type* are both simply stored as "house").

The data is split on the first night of the first lockdown into *Prec* (pre-covid) and *Post*. *Prec* now contains c. 9m instances used to train and validate, *Post* has c. 8m instances used to predict the impact (a visualization can be found in Appendix 1.2). For the booking model development, only the data of the latest 400 nights is used for training because only that share of the data has sufficient bookings (Appendix 1.3).

Cancellation and delisted items required for the loss deconstruction from research question Q3 are also identified in step I of the framework (*Figure 2*). The cancellations in night *n* are defined as available listings that were unavailable for the same night *n* in the previous update. Therefore, cancellations are accounted for at the night *n* of the actual loss, not at the point in time the cancellation was made. This approach to modeling cancellations is different to the approach used by Hu & Lee, who use the contents of review messages to identify cancelled listings (Hu und Lee 2020).

While cancellations can simply be identified from the booking states of the listings in the data, delisted items (those taken out of the system by the hosts) are more complex, since their data is entirely unavailable in the set. A two-step process enables estimating delisting loss: First, in a separate data-frame called *Delistings*, all theoretically possible but missing *Post* (all data after the first lockdown) combinations of listings *l* and nights *n* are synthesized using the latest available data for each *l*. Therefore, synthesizing in this case describes the process of using a listing's data from available nights to recreate the listing's missing nights. Together with *Post*, *Delistings* now contains $L_{Post} * N_{Post}$ combinations (i.e., instances for all listings across all nights after the first lockdown). Delisting items is not exclusively caused by C19 and is observed frequently in *Prec* (c. 25,000 listings are in the system per day in *Prec*, while c. 35,000 listing ids are in the system). Keeping all synthesized instances would imply that in absence of C19 no items would have been delisted, which significantly overstates the hypothetical *Post* supply market. Therefore, in the second step, per *n* in *Post*, the missing listings are randomly added from

Delistings to *Post* only to the point where the latter's count of *l* at night *n* matches the median count per *n* in *Prec*, which perpetuates the implicit delisting rate observed from *Prec.* For the calculation of the median listing count in *Prec*, only data from the most recent six months is used to capture the year's market growth. A visualization of the synthesizing process can be found in Appendix 1.4, the respective items are denoted as *Taken_Offline* in the code. The final data set as fed into the prediction pipeline now contains 16.7m instances, of which synthesized items make up 6.3%.

In step II of the framework, several base models are trained on randomly sampled and differently sized sets of training instances (10k, 50k, 100k, 500k, 1m, 1.5m) to observe how the different algorithms perform at scale (computational power is limited, the entire analysis is run on CPU with 16GB RAM). The base models tested include regressors and classifiers from Python libraries LGBM, sklearn's LassoCV, DecisionTrees, RandomForests, AdaBoost and XGBoost (Appendix 10). Since prices are continuous and the booking dummy is discrete ({1, 0}), price estimation is a regression exercise and booking prediction a classification task. Different metrics for performance evaluation are therefore required. The regression model is optimized towards commonly used mean absolute error (MAE), which is defined in Eq.IV (*Figure 4*). Throughout the supplementary analysis, most pricing results are presented as median values. Since the final model is applied to data that contains previously seen instances at different points in time, the expectation prevails that unregularized models which fit closely to the training data yield higher validation performance. Although the cost of misclassification is the same for false positives (F_p) and false negatives (F_n), choosing model accuracy as the core classification metric while evaluating a closely fitting model on an imbalanced dataset might lead to misguiding results. If, for instance, the base rate of bookings would be 5%, a worthless model predicting 100% vacancy still achieved 95% accuracy. *F1-score* (Eq.V.), a metric that considers true

positives with respect to both predicted and observed positives, is therefore used as the key classification metric (*Figure 4*). All metrics used are common practice in machine learning.

***Figure 4**:* Selected metrics to measure model performance

Equation IV. $$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Equation V. $$Recall = \frac{T_p}{T_p + T_n}, \quad Precision = \frac{T_p}{T_p + F_p}, \quad F1 = 2 * \frac{precision * recall}{precision + recall}$$

Source: Author

Base models whose performance justifies computational scaling are qualitatively selected for hyperparameter (HP) tuning in step III of the framework. HP tuning identifies the set of parameters applied to the algorithm itself, e.g., learning rates, depths and regularization.

Model and HP performance is most commonly validated using an approach called *k-fold Cross Validation* (Anguita 2012). This approach randomly samples a set of *k* folds, iteratively trains on *k*-1 folds and validates on the residual fold, until all folds have inherited the role of validating once. In time-series modeling, however, projections are made for values yet to be observed, hence there prevails a clear chronology in training and predicting. Using random sampling in time-series implies that theoretically still unobserved values would be used to train the models, which causes the training set to contain values and information that the models do not have access to during deployment (Brownlee 2020). Therefore, in step III of the framework, a forward-chaining validation technique called *Nested Cross-Validation* (Cochrane 2018) is applied to make sure chronological data leakage is avoided. Although the data is still split into folds like with the commonly used *k-fold Cross Validation*, the nested approach makes sure that validation data does not appear before the training data and that the model makes projections exclusively for the future and not for the past: For each *k*, the model is trained on all folds up to *k-1* and validated on fold *k*. While solving the data leakage problem, this approach also causes the training data to grow with each fold. Since some models might scale well with increasing

amounts of data, the performance of the earlier folds needs to be interpreted with care: The goal of step III is to identify the best-performing algorithms and whether or not their performance is increased with adjusted hyperparameters. Depending on how quickly the models' base algorithms improve in performing at scale in step II, validation metrics are calculated only from the latest 3 folds because those are trained with significantly more data than the first folds.

In addition to the obtained machine learning models, the daily aggregations of all listings (medians for *P*, sums for *B* and *R*) are compared to the *Prec* monthly medians in the respective month. For example, the total revenues (across all l in L) on June 20th, 2020 are compared to the median total daily revenues in June 2019. While the combination of the previously discussed machine learning models (for prices and bookings) is referred to as *ML_models*, this alternative prediction method is denoted as *Heuristic* and yields a sanity check for the machine learning predictions. Note that as opposed to the machine learning approach, *Heuristic* does not identify discrepancies at listing level, but compares aggregations over time.

Finally, the best-performing models for each task are used to add $\hat{P}_{hyp_{l,n}}$ and $\hat{B}_{hyp_{l,n}}$ to the data in step IV and to compute the estimated loss $\widehat{D}_{L,n}$ in step V. Having fitted the models with feature sets *X1* and *X2*, parameter weights (e.g., *feature_importances_* (FI)) are analyzed in step VI to identify forces driving $\widehat{D}_{L,n}$ and explore further considerations. In addition to the models' feature weights, the identifiers created for cancellations and delisted items during the curation process are used as dimensions in step VI. The findings of the resulting dimensional analyses are descriptive to show the developed models' capacity to dive into the dynamics of loss.

In order to compute lost fee payments to the entity of the Lisbon city government as part of the case study, a nightly fee of 2€ (The Portugal News 2020) can be multiplied with all lost bookings (those whose hypothetical booking Boolean is 1 and its observed one is 0). In reality, this 2€ fee applies only to stays shorter than 7 nights, yet this is neglected since there is no way to identify the visitor via *Inside Airbnb* data.

4. Results

4.1. Findings in model development

Base algorithm selection: All base model selection results can be found in Appendix 2. For the model selection in both pricing and bookings, LGBM, LassoCV and DecisionTree compute much faster than the remaining algorithms while still being able to increase their performance at scale (e.g., fitting the RandomForestRegressor from sklearn at 1.5m samples takes >25 minutes, while LGBM's LGBMRegressor trains for the same task in few seconds). The three algorithms are therefore selected for hyperparameter tuning.

Hyperparameter tuning: All results from the forward-chaining HP validation can be found in Appendix 3. As explained in the methodology section, observing well-scaling models indicates that only the latest 3 folds are used for HP tuning, as those models' performance only unlocks if the models are trained using many instances. The automated model selection process identifies a DecisionTreeRegressor with validation MAE of 23.7$ as the pricing model (Appendix 3.1) and also a DecisionTreeClassifier with a validation F1-score of 64.4% as the booking model (Appendix 3.2). All other models have higher MAEs (price) or lower F1-scores (bookings) across the folds. Both final models have empty hyperparameter dictionaries. As expected, the fact that most validation instances are represented in the training data (at different points in time) causes these strongly fitting models to still regularize well on validation data.

It should also be mentioned that far more combinations of hyperparameters have been tested than indicated in the Appendix, although many of them would not converge on CPU for the implied computational expense inherent with their impact on the base algorithm's workload.

Identifying relevant dimensions: To the pricing- and booking model, the features [*latitude, longitude, maximum_nights, accommodates, bathroom*] and [*DoY*, *latitude*, *longitude*, *number_of_reviews*, *Weekday*] are most decisive according to feature importance, respectively (Appendix 4). This implies that the size of a listing (via *accommodates* and *bathrooms*) is mostly

relevant to its price, the point in time (via DoY and Weekday) is mostly relevant to bookings and location (via *latitude* and *longitude*) is highly important to both bookings and prices.

As noted in the methodology, the bottom-up approach of using listing-level machine learning for the predictions enables further dimensional analysis. As they prove to be most relevant to supply and demand, the dimensions location (*by location*), size (*by size*) and timing (*by time*) are analyzed together with the impact of delisted items and cancellations (*by forces*) in more detail in section 4.2.

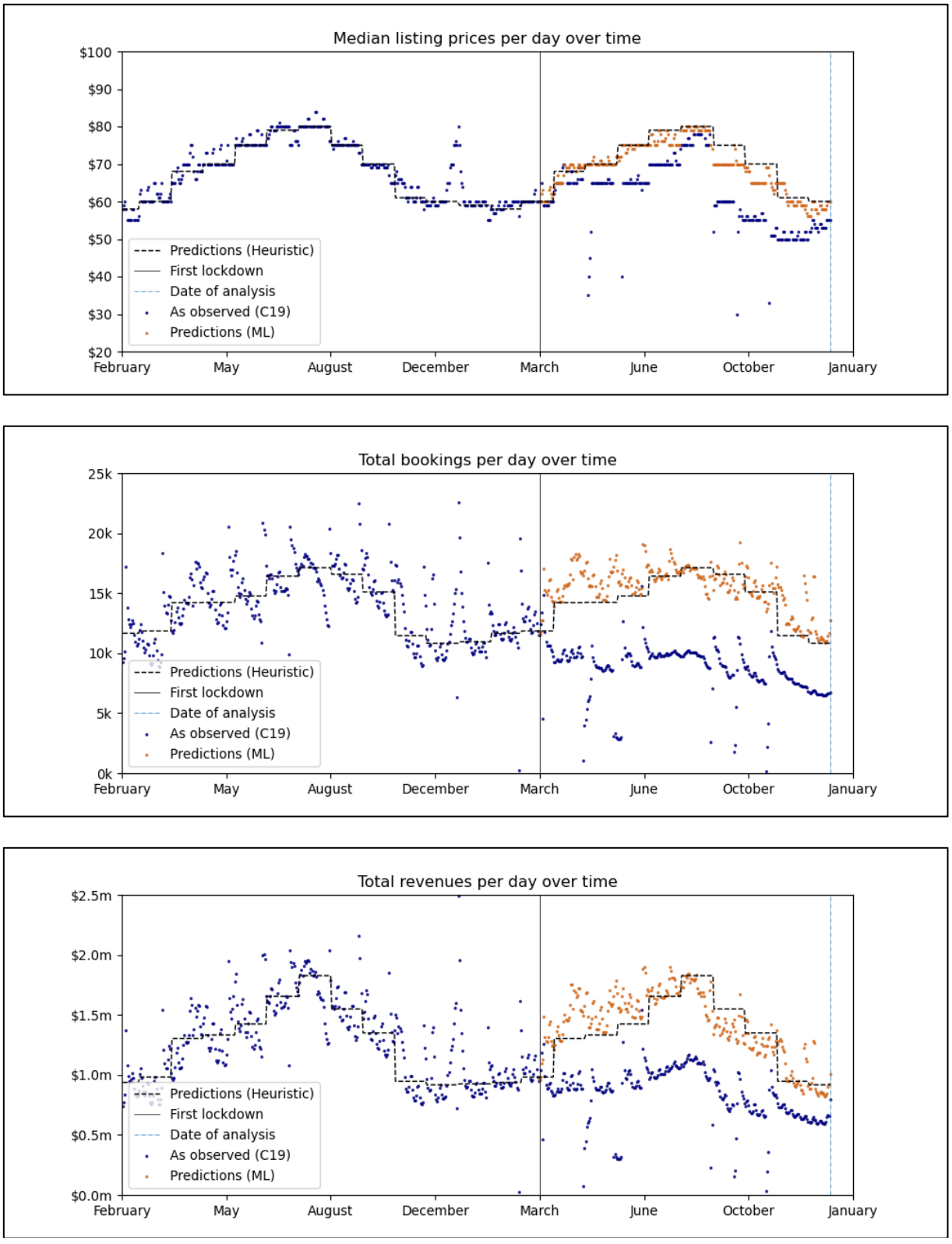
4.2. Findings in model deployment

For the sake of contextual comprehensibility, the identified dimensions are presented in the following order: *by time*, *by forces*, *by location* and *by size*. Just like the findings from the previous section, all findings from model deployment are visualized in the Appendix, examples of which are shown in this section. All figures (except frameworks, e.g., *Figure 2*) are fully and automatically generated by the Python3 code (including maps, whose background images are created using *openstreetmap* (OSMF 2020).

Appendix 5.1 (*Figure 5* shows it as an example) shows the daily predictions of the machine learning models (maroon scatter) and heuristic (black dotted) in direct comparison with actual observations (blue scatter). The figures visualize how significantly different the Lisbon observations have been distributed ever since the night of the first lockdown and how seasonality in prices, bookings and revenues has almost entirely vanished. The chart on median listing prices per day over time shows how median prices have decreased to prices usually expected in November by August 2020 and how they have since dropped to all-time lows. The daily total bookings indicate that with relatively little variation, between 5,000 and 10,000 bookings have been missing in Lisbon each day since the first lockdown. Revenues have dropped by roughly $0.5 million to $1 million per day.

4.2.1. Impact by time

***Figure 5**:* Comparison of observed and predicted values per night



Source: Author

In order to better identify the seasonality of the losses, Appendix 5.2-5.4 aggregate the results by months: Median listing prices were down by roughly 10$ throughout June, rebounded in August and then dropped 15$ below the predictions during the second wave in October (Appendix 5.2). This proves that the supply market has significantly lost power over pricing. Bookings in Lisbon are found to have tumbled by roughly a quarter million each month (Appendix 5.3). In monetary value, this translates into c. $20m in revenue losses each month (Appendix 5.4).

From a cumulative perspective, since the beginning of the first lockdown on March 18 (Appendix 5.5), Lisbon is predicted to have lost $140m and $160m by *Heuristic* and *Ml_models*, respectively. The loss stated by *Heuristic* is slightly lower because less listings were available in the beginning of 2019 (Appendix 1.4), which decreases the model's predicted hypothetical revenues. So far, the cumulative loss in the city of Lisbon is concluded to be roughly $150m. Applying the currency-adjusted 2€ nightly city government fee to the lost bookings, more than $4m have been lost in government fee payments alone (Appendix 5.5). Remember that the government fees are referring exclusively to the tourist fees, missing tax payments are not reflected.

4.2.2. Impact by forces

Mapping the cumulative losses to the identifiers of cancellations and delisted items, their relative shares in overall loss can be visualized (Appendix 6), which unveils the three forces that ultimately drive loss: *Cancellation loss*, *Delisting loss* and *Aversion loss*. The latter denotes the residual cumulative loss predicted by *ML_models* after *Cancellation loss* and *Delisting loss*. It is called *Aversion loss* because it is captured in neither of the other loss forces and therefore is assumed to be rooted in consumer aversion to book listings. Note that cancellations are calculated only from listings that have originally been in the system, hence the theoretical cancellation losses for all items that have been taken offline because of C19 is embedded in *Delisting*

loss. The upper figure in Appendix 6 shows that so far, cancellations, delisted items and aversion have contributed $20m, $60m and $80m to overall cumulative loss, respectively. As expected, the share of *Cancellation loss* has been significantly higher shortly after the first lockdown, which can be derived from the lower figure in Appendix 6. Since its peak in April, *Cancellation loss* has decreased its share in overall losses from 80% to 20%. The observation that *Cancellation loss* has reached its peak only weeks after the first lockdown is induced by the fact that it accounts for actual losses, which occur on the hosting night, not the cancellation night (bookings are made in advance). *Delistings loss* is found to capture roughly 40% of cumulative loss. As previously explained, bookings that are not made in the first place (neither cancelled nor delisted) are embedded in *Aversion loss*. The losses assigned to this force have become visible 2 months after the first lockdown and currently make up c. 40% of cumulative hosting revenue loss.

4.2.3. Impact by location

Applying the model predictions to the dimension of location demonstrates how significantly different C19 has impacted neighborhoods (Appendix 7). For instance, aggregating loss by neighborhood shows that besides the central Lisbon area, especially Lourinha, Santo Isodoro, Ericeira, Sintra and Cascais have been hit hardest (Appendix 7.1). Prices have dropped the most in denser areas, which can be interpreted from the distribution of colors in the lower right map in Appendix 7.1: Green and blue values, which denote higher discounts, are mostly observed in central Lisbon. Taking an isolated look at the impact in central Lisbon (Appendix 7.2) reveals that half of the overall Lisbon area loss is concentrated in only two neighborhoods: Santa Maria Maior (lower Alfama) has lost $48m and Misericórdia (Bairro Alto, Baixa, Chiado) has lost $38m. The five hardest-hit districts also contain Arroios ($22m), Santo Antonio ($17m) and São Vicente ($15m). Santo Antonio comprises Marquês and Avenida, São Vicente is like Santa Maria Maior a part of Alfama.

Appendix 7.3 shows how prices have changed in central Lisbon. Just like with the results from Lisbon area (Appendix 7.1), it can be observed that larger decreases in prices (brighter values) can be observed in the more central areas, i.e., those that have also seen the highest hosting revenue losses. This indicates that while most areas have been hit by missing bookings, these central areas have additionally seen strong decreases in prices and consequently lost more money. One explanation for this dynamic might be that consumers are more elastic to price when two listings are closer to each other and substitution is more effortless, implying that the combination of C19 and a high density of listings has spurred a 'price war' in Lisbon's touristic areas.

4.2.4. Impact by size

As with the other dimensions, the bottom-up modeling approach also allows to group losses by listing size. As an example for supplementary analysis, larger and smaller listings are split at 3 bedrooms to identify differences in means (Appendix 7.8). Occupancy rates for larger houses have declined less (-11%) than small listings (-14%) since the first lockdown, which is in line with findings from Hu et al. Furthermore, the predictions by *ML_models* indicate that while mean prices of smaller listings have remained unchanged (means are not robust to outliers, which is why they do not detect the losses reflected in median prices), the mean prices for larger houses have gained 7%. Despite the larger absolute mean loss per night for large listings ($8 as opposed to $6), the much higher base prices inherent with more spacious listings cause large listings' relative revenue loss (-6%) to be smaller than the one measured across smaller listings (-13%). These results are exemplary and descriptive, i.e., no causal relationships are tested.

5. Discussion

5.1. Theoretical and practical contributions

The research objective of answering questions Q1-Q5 has been accomplished in this paper. For Q1, quantified financial loss estimations and a breakdown on prices and bookings has been supplied. The ability to recursively model hosting revenue loss in the short-term rental market for Q2 has been achieved in the case study and can be updated at any time by running the Python script again. Since bookings have been optimized towards F1-score, which takes into account both precision and recall, the decision thresholds for bookings can be interpreted as balanced and the booking predictions as neither aggressive nor conservative. Still, with regard to the fact that bookings steadily increased in 2019 (Appendix 1.4) but were synthesized at late 2019 levels, it might be reasonable to assume that the loss estimates are rather conservative. Evidently, this only holds under the assumption that even more listings would have surfaced on Lisbon's 2020 supply market in absence of C19.

Inherent with the bottom-up methodology, predictions across listings and nights can be deconstructed for further analysis, which is highly important for the research objective behind Q3: For example, being able to breakdown losses into aversion, cancellation and delisted items shows that the progressive side-effects of long-term renting discussed by Dolnicar and Zare account for a maximum of 40% of cumulative losses. It has to be taken into account that this includes the extra cancellations that would have been made in absence of the delisting wave. Also, their socio-economic interpretation is based on the disputable assumption that all money lost from delisting flows into long-term renting, hence the actual share realized for long-term renting is likely to be even less than 40%. This implies that at least the other 60% of the cumulative losses, i.e., roughly $100m still have been lost in Lisbon without any positive side effect absorbing them.

Q4, the quest for features that drive prices and bookings, has also been successfully completed. In the code, the entire feature sets are ranked by importance, only the most relevant are displayed in this paper (Appendix 4). The use cases of identifying relevant features have been presented among the exemplary dimensions of time, location and size. In Q5, mapping the predictions to listing coordinates and neighborhoods revealed further insights into the hardest-hit districts. Losses are distributed in a way that leaves the two neighborhoods of Santa Maria Maior and Misericórdia losing as much money as all other districts in Lisbon area combined, which is an interesting revelation and should be considered by the government when injecting cash into the local economy.

Apart from being able to apply well to the case study city of Lisbon, the value of this paper lies in those analyses that have not yet been done. The way for these has been paved by coding extensive curation and model development infrastructure. The Python3 script that runs the entire analysis comprises 1,133 lines of relatively information-dense code. Up to the start of the supplementary case study analysis, which is inherently bound to the city of Lisbon (e.g., 2€ city government fee, maps with fixed coordinates), the entire pipeline can be applied to other cities by simply changing the identifiers for “city”, “country”, and “start_date_pandemic”, which denotes the start of the first lockdown. Proof that the modeling approach generalizes well to other cities can be found in Appendix 7.9, which shows exemplary figures from the same script being run on data from Venice, Italy.

In a nutshell, the tools developed in this paper help to quantify C19’s severe impact on short-term renting via Airbnb and unlock significant potential for global analysis, targeting any city whose data is available on *Inside Airbnb*.

5.2. Limitations and further research

As indicated in the previous paragraph, most value supplied by this paper lies in its own limitation of scope. The largest share of enabled analysis is still to be done on many more cities, longer time periods and broader sets of features. While the obtained predictions could be applied in many ways, some examples are listed below:

- The impact on pricing could be analyzed across neighborhood aggregations
- Cumulative loss per neighborhood could be mapped over time or aggregated by month
- Dimensional analysis has been done on exemplary basis with only three of the most relevant dimensions (time, location and size). Much more analysis can be done on additional features (e.g., "*review_score_rating*" could be used to visualize impact of ratings on losses)
- Losses in government payments could be further inspected by multiplying the identified losses with income tax rates
- The losses over time could be compared to the infection rates over time. Especially when applied across cities with different lockdown behaviors and similar infection rates, the costs of excess stringency could be quantified

The findings from the dimensional analyses are exemplary and descriptive to show the depth of analysis enabled by modeling losses through a bottom-up approach. Although significance of findings can be assumed given the vast sample sizes compared (*Prec* and *Post* contain almost 17m instances), further research could measure actual relationships using significance levels.

As indicated in the methodology section, computational expense has significantly limited the research depth of this paper. Running the script for a city that has as many listings as Lisbon requires multiple hours of computation, the Lisbon directory with all curated data, workflow-related pickle files and models requires roughly 24GB of disk space. Further research should

utilize GPUs (Graphics Processing Units) to put the bottom-up modeling approach developed in this paper to a test with broader cities and even consider developing a more holistic model across the entire *Inside Airbnb* dataset. Furthermore, deep learning has yet to work its way into quantifying C19 impact. Future research could establish similar approaches using neural networks.

The pipeline that synthesizes delisted items still has intense workloads, since the *Post* dataframe needs to be filtered and sorted for millions of items given their night and listing id. A future model might be able to simply predict the feature arrays for these synthesized instances using e.g., naïve bayes to eradicate lengthy computations. Furthermore, perpetuating the delisting rate from *Prec* data by random sampling is blind to the non-random occurrence of delisted items. The probability of a listing to be in the system assumably depends on whether the listing had been in the system on the previous day. Specifically, delisted items are likely to remain delisted. Randomly adding instances has been selected because of its simplicity, but more advanced selection methods could be applied in future research.

Finally, despite *Airbnb* having large market-shares around the world, the short-term renting market is impacted by C19 not just through the dynamics of a single company. Future research might focus on building more integrative models that feed data with various sources.

6. Bibliography

WHO. 2020. *WHO Coronavirus Disease (COVID-19) Dashboard.* 10 23. https://covid19.who.int.

Callaway. 2020. "COVID and 2020: An extraordinary year for science." *nature.*

Junxiong Li, Thi Hong Hai Nguyen, J. Andres Coca-Stefaniak∗. 2020. "Coronavirus impacts on post-pandemic planned travel behaviours." *Annals of Tourism Research.*

Karabulut, Gokhan , Mehmet Huseyin Bilgin, Ender Demir, and Asli Cansin Doke. 2020. "How pandemics affect tourism: International evidence T." *Annals of Tourism Research.*

Abhinav Sharma, Juan Luis Nicolau. 2020. "An open market valuation of the effects of COVID-19 on the travel T and tourism industry." *Annals of Tourism Research.*

Hossain, Mokter. 2020. "The effect of the Covid-19 on sharing economy activities." *Journal of Cleaner Production.*

Qiu, Richard T.R. , Jinah Park, ShiNa Li, and Haiyan Song. 2020. "Social costs of tourism during the COVID-19 pandemic." *Annals of Tourism Research.*

Kock, Florian, Astrid Nørfelt, Alexander Josiassen, A. George Assaf, and Mike G. Tsionas. 2020. "Understanding the COVID-19 tourist psyche: The Evolutionary Tourism Paradigm." *Annals of Tourism Research.*

Dolnicar, Sara, and Samira Zare. 2020. "COVID19 and Airbnb–Disrupting the disruptor." *Annals of Tourism Research.*

ANMP. 2020. *Associacao Nacional Municipios Portugueses.* 10 3. https://www.anmp.pt.

Ginzel, Leon. 2019. "Beste Lage, zu viele Touristen." *Spiegel*, 5 3.

AirDNA. 2020. *MarketMinder Lisbon.* 10 10. https://www.airdna.co/vacation-rental-data/app/pt/lisboa/lisbon/overview.

Reuters. 2020. *Portugal registers first two cases of coronavirus: health minister.* 10 02. https://www.reuters.com/article/us-health-coronavirus-portugal-idUSKBN20P1BB.

—. 2020. *Reuters.* 10 3. https://www.reuters.com/article/us-health-coronavirus-portugal-tourism/coronavirus-fears-pressure-portugals-tourism-dependent-economy-idUSKBN20X1V0.

—. 2020. *Reuters.* 10 3. https://www.reuters.com/article/us-health-coronavirus-portugal-idUSKBN20Z3OP.

—. 2020. *UPDATE 2-Tourism between Spain, Portugal to be suspended due to coronavirus.* 10 2. https://www.reuters.com/article/health-coronavirus-portugal-spain-idUSL8N2B8194.

Portuguese American Journal. 2020. *Portuguese American Journal.* 10 3. https://portuguese-american-journal.com/pandemic-portugal-to-declare-state-of-emergency-update/.

John Hopkins University. 2020. *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.* 10 3. https://github.com/CSSEGISandData/COVID-19.

Cox, Murray. 2020. *Get the data.* 11 06. http://insideairbnb.com/get-the-data.html.

Cochrane, Courtney. 2018. *Time Series Nested Cross-Validation.* May 19. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9.

The Portugal News. 2020. *The Portugal News.* 10 25. https://www.theportugalnews.com/news/lisbon-to-double-tourist-tax-from-next-year/47192.

OSMF. 2020. *OpenStreetMap Foundation.* 10 30. https://www.openstreetmap.org/copyright.

Rocca, Rosa Anna La. 2015. "Tourism and Mobility. Best Practices and Conditions to Improve Urban Livability." *Journal of Land Use, Mobility and Environment* 311-330.

Anguita, Davide. 2012. "The 'K' in K-fold Cross Validation." *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.* Bruges: University of Genova.

Brownlee, Jason. 2020. *Data Leakage in Machine Learning.* 12 02. https://machinelearningmastery.com/data-leakage-machine-learning/.

Harvard Business Review. 2019. "Research: When Airbnb Listings in a City Increase, So Do Rent Prices." *Harvard Business Review.*

Brownlee, Jason. 2020. *Ordinal and One-Hot Encodings for Categorical Data.* 08 17. Accessed 08 20, 2020. https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/.

Warren, Hayley. 2020. "Airbnb Hosts Resist Lisbon's Plan to Free Up Housing." *Bloomberg Business*, 7 28.

Airbnb. 2020. *A Message from Co-Founder and CEO Brian Chesky.* 5 5. Accessed 10 10, 2020. https://news.airbnb.com/a-message-from-co-founder-and-ceo-brian-chesky/.

Chen, Guangwu , Mingming Cheng, Deborah Edwards, and Lixiao Xu. 2020. *COVID-19 pandemic exposes the vulnerability of the sharing economy* . Sydney.

Yang, Yang, Hongru Zhang, Xiang Chen, and Y. 2020. "Coronavirus pandemic and tourism: Dynamic stochastic general equilibrium modeling of infectious disease outbreak." *Annals of Tourism Research.*

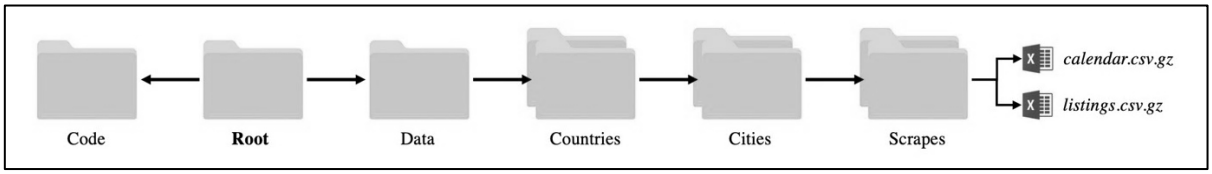
Wisniewska, Aleksandra. 2019. *Are Airbnb investors destroying Europe's cultural capitals?* Lisbon, 9 5.

Hu, Maggie., and Adrian D Lee. 2020. *Airbnb, COVID-19 Risk and Lockdowns: Local and Global Evidence.* Deakin University.

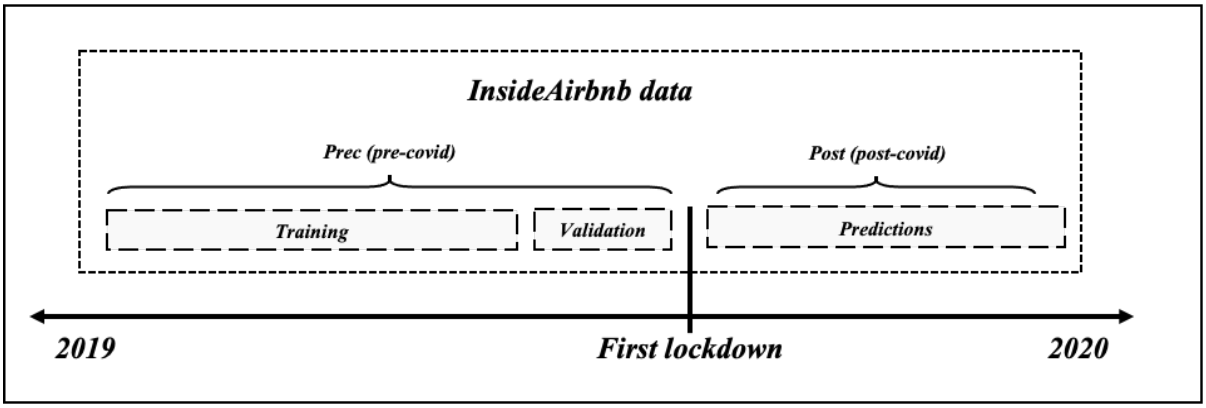
Appendix

Appendix 1. The data

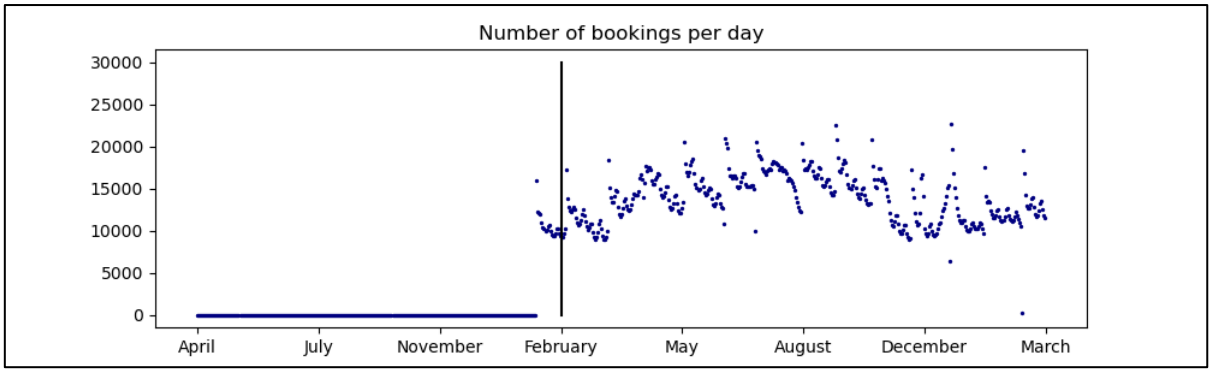
Appendix 1.1. Directory structure



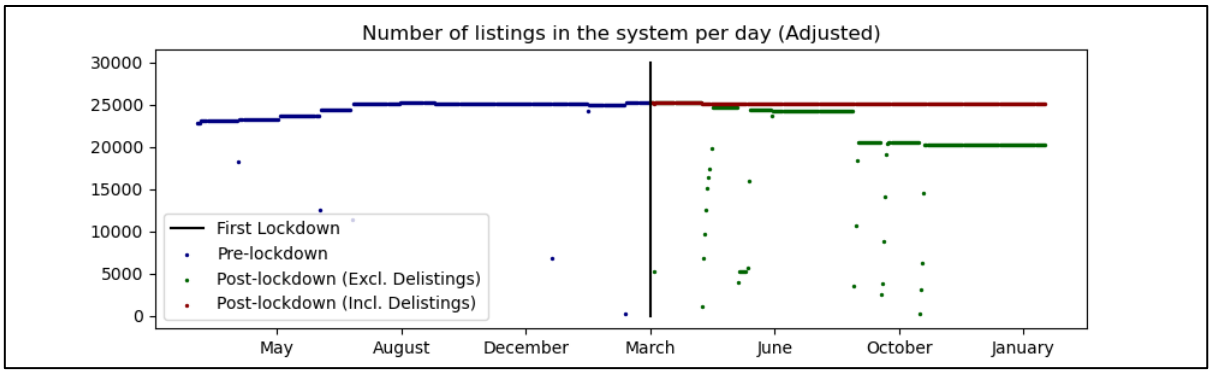
Appendix 1.2. Allocation of data to training and predictions



Appendix 1.3. Availability of booking data over time



Appendix 1.4. Synthesizing delisted items



Appendix 1.5. Data dictionary of features used

Feature	Types	Missing (%)	# diff. values	Most common
listing_id	int64	0.00	24235	10362958
date	object	0.00	366	2021-05-03
available	object	0.00	2	t
price	object	0.06	1011	$60.00
minimum_nights	float64	0.01	67	2
maximum_nights	float64	0.01	195	1125
id	int64	0.00	24235	985087
host_response_time	object	31.45	4	within an hour
host_response_rate	object	31.45	59	100%
host_acceptance_rate	object	11.02	85	100%
host_is_superhost	object	0.00	2	f
neighbourhood_cleansed	object	0.00	128	Santa Maria Maior
latitude	float64	0.00	9477	38.7119
longitude	float64	0.00	11196	-9.1361
property_type	object	0.00	39	Apartment
room_type	object	0.00	4	Entire home/apt
Accommodates	int64	0.00	20	2
Bathrooms	float64	0.11	29	1
Bedrooms	float64	0.05	17	1
Beds	float64	0.58	34	1
bed_type	object	0.00	5	Real Bed
security_deposit	object	23.87	129	$0.00
cleaning_fee	object	18.38	129	$0.00
guests_included	int64	0.00	18	1

Feature	Types	Missing (%)	# diff. values	Most common
extra_people	object	0.00	68	$0.00
number_of_reviews	int64	0.00	435	0
review_scores_rating	float64	18.89	53	100
review_scores_accuracy	float64	18.95	9	10
review_scores_cleanliness	float64	18.93	9	10
review_scores_checkin	float64	18.98	9	10
review_scores_communication	float64	18.96	8	10
review_scores_location	float64	18.97	9	10
review_scores_value	float64	18.97	9	9
instant_bookable	object	0.00	2	t
is_business_travel_ready	object	0.00	1	f
cancellation_policy	object	0.00	8	strict_14_with_grace_period
require_guest_profile_picture	object	0.00	2	f
require_guest_phone_verification	object	0.00	2	f

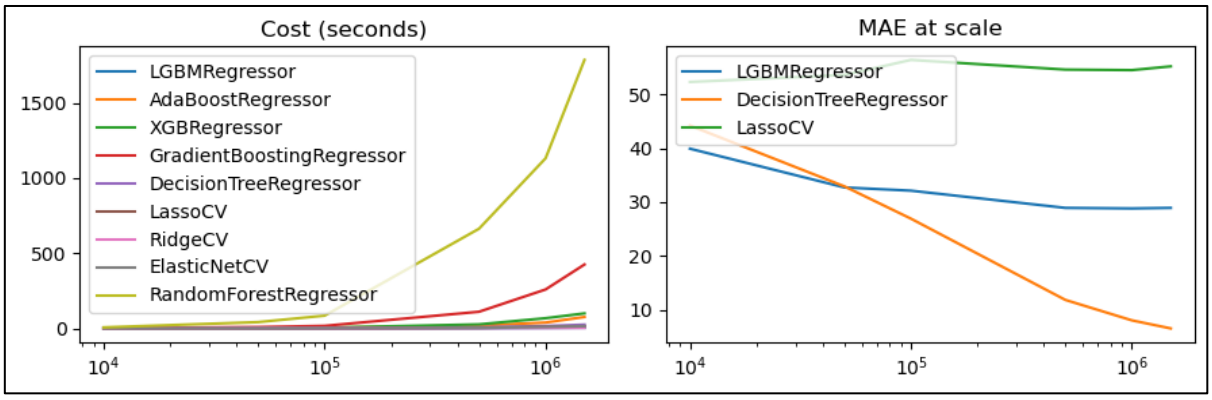
Appendix 2. Base model selection

Appendix 2.1. Performance at scale: Pricing model base algorithms

Instances	Model	RMSE	MAE	R2	Cost (s)
50000	LGBMRegressor	69.3	32.7	0.51	0
50000	AdaBoostRegressor	134.5	101.5	-0.85	4
50000	XGBRegressor	75.4	30.6	0.42	3
50000	GradientBoostingRegressor	67.0	34.1	0.54	9
50000	DecisionTreeRegressor	100.2	32.9	-0.03	1
50000	LassoCV	99.0	53.6	-0.0	0
50000	RidgeCV	17585.4	693.6	-31544.48	0
50000	ElasticNetCV	99.0	53.6	-0.0	0
50000	RandomForestRegressor	66.6	26.3	0.55	42
100000	LGBMRegressor	128.3	32.1	0.4	0
100000	AdaBoostRegressor	172.1	81.4	-0.08	4
100000	XGBRegressor	124.5	30.7	0.43	5
100000	GradientBoostingRegressor	142.0	36.5	0.26	17
100000	DecisionTreeRegressor	121.6	26.9	0.46	1

Instances	Model	RMSE	MAE	R2	Cost (s)
100000	LassoCV	165.4	56.4	0.0	1
100000	RidgeCV	13507.4	394.7	-6661.63	0
100000	ElasticNetCV	165.4	56.4	0.0	1
100000	RandomForestRegressor	115.4	23.0	0.51	85
500000	LGBMRegressor	64.1	28.9	0.74	2
500000	AdaBoostRegressor	135.3	65.9	-0.17	19
500000	XGBRegressor	60.1	27.0	0.77	27
500000	GradientBoostingRegressor	86.1	35.1	0.53	111
500000	DecisionTreeRegressor	55.8	11.8	0.8	7
500000	LassoCV	125.2	54.6	0.0	4
500000	RidgeCV	710700.4	18844.3	-32223914.08	2
500000	ElasticNetCV	125.2	54.6	0.0	4
500000	RandomForestRegressor	37.7	10.9	0.91	663
1000000	LGBMRegressor	79.0	28.8	0.67	4
1000000	AdaBoostRegressor	137.8	64.2	0.0	39
1000000	XGBRegressor	72.2	27.0	0.73	68
1000000	GradientBoostingRegressor	111.1	35.5	0.35	259
1000000	DecisionTreeRegressor	40.4	8.0	0.91	16
1000000	LassoCV	137.8	54.5	0.0	9
1000000	RidgeCV	12957847.3	369972.9	-8832472267	3
1000000	ElasticNetCV	137.8	54.5	0.0	10
1000000	RandomForestRegressor	46.4	8.0	0.89	1131
1500000	LGBMRegressor	76.8	28.9	0.75	5
1500000	AdaBoostRegressor	143.5	63.8	0.14	76
1500000	XGBRegressor	65.1	26.9	0.82	100
1500000	GradientBoostingRegressor	129.9	36.0	0.29	425
1500000	DecisionTreeRegressor	41.6	6.5	0.93	25
1500000	LassoCV	154.4	55.2	0.0	14
1500000	RidgeCV	2437051.5	70547.6	-248978425.34	4
1500000	ElasticNetCV	154.4	55.2	0.0	15
1500000	RandomForestRegressor	37.2	6.6	0.94	1785

Appendix 2.2. Evaluating pricing base algorithm performance

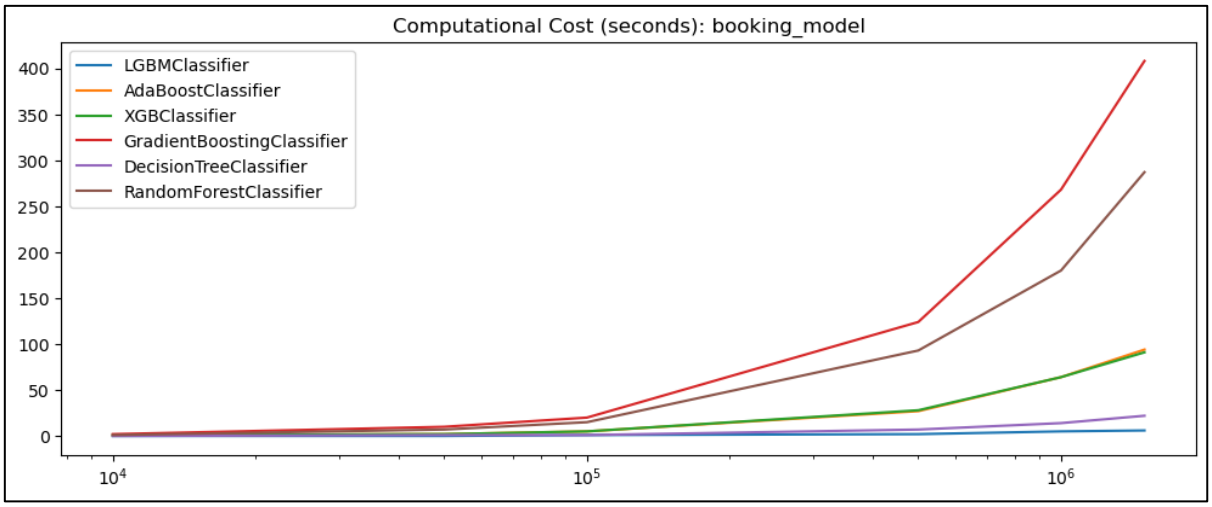


Appendix 2.3. Performance at scale: Booking model base algorithms

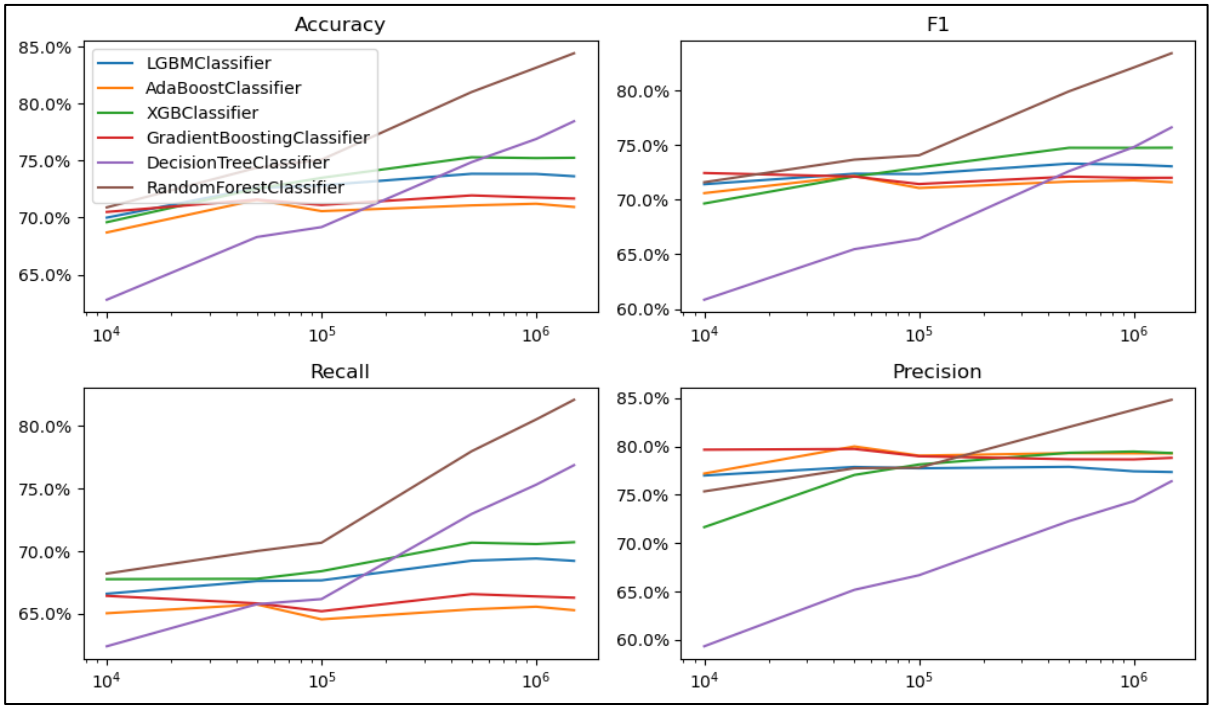
Instances	Model	Accuracy	Recall	Precision	F1	ROC	Cost (s)
50000	LGBMClassifier	0.726000	0.676205	0.778838	0.723902	0.729314	0
50000	AdaBoostClassifier	0.715600	0.657520	0.800087	0.721831	0.723701	2
50000	XGBClassifier	0.725400	0.677986	0.770598	0.721331	0.727812	2
50000	GradientBoostingClassifier	0.715800	0.658432	0.797485	0.721318	0.723416	10
50000	DecisionTreeClassifier	0.683000	0.657768	0.651778	0.654759	0.681002	1
50000	RandomForestClassifier	0.743800	0.700117	0.777537	0.736799	0.744893	7
100000	LGBMClassifier	0.728300	0.676750	0.777486	0.723629	0.731082	1
100000	AdaBoostClassifier	0.705700	0.645662	0.790601	0.710819	0.713918	5
100000	XGBClassifier	0.734900	0.684080	0.781421	0.729517	0.737306	5
100000	GradientBoostingClassifier	0.711100	0.652111	0.789945	0.714441	0.718272	20
100000	DecisionTreeClassifier	0.691700	0.661822	0.666885	0.664344	0.689538	1
100000	RandomForestClassifier	0.750800	0.706770	0.778142	0.740741	0.751128	15
500000	LGBMClassifier	0.738460	0.692426	0.778987	0.733161	0.740268	2
500000	AdaBoostClassifier	0.710760	0.653659	0.793166	0.716687	0.718499	27
500000	XGBClassifier	0.752940	0.706809	0.793513	0.747656	0.754645	28
500000	GradientBoostingClassifier	0.719480	0.665774	0.786792	0.721242	0.724802	124
500000	DecisionTreeClassifier	0.748660	0.729680	0.722877	0.726263	0.747155	7
500000	RandomForestClassifier	0.810160	0.779678	0.820180	0.799417	0.809284	93
1000000	LGBMClassifier	0.738320	0.694179	0.774377	0.732088	0.739689	5
1000000	AdaBoostClassifier	0.712190	0.655729	0.792939	0.717836	0.719644	64
1000000	XGBClassifier	0.752220	0.705692	0.794802	0.747601	0.754159	64
1000000	GradientBoostingClassifier	0.717670	0.663943	0.786680	0.720119	0.723251	268
1000000	DecisionTreeClassifier	0.769000	0.753038	0.743513	0.748245	0.767705	14

Instances	Model	Accuracy	Recall	Precision	F1	ROC	Cost (s)
1000000	RandomForestClassifier	0.831520	0.805069	0.837990	0.821200	0.830531	180
1500000	LGBMClassifier	0.736293	0.692271	0.773611	0.730684	0.737820	6
1500000	AdaBoostClassifier	0.709373	0.652974	0.792901	0.716166	0.717285	94
1500000	XGBClassifier	0.752493	0.707171	0.793218	0.747727	0.754253	91
1500000	GradientBoostingClassifier	0.716747	0.662920	0.788273	0.720183	0.722709	408
1500000	DecisionTreeClassifier	0.784533	0.768587	0.764111	0.766342	0.783345	22
1500000	RandomForestClassifier	0.844160	0.820713	0.848305	0.834281	0.843170	287

Appendix 2.4. Booking model base algorithms computational expense



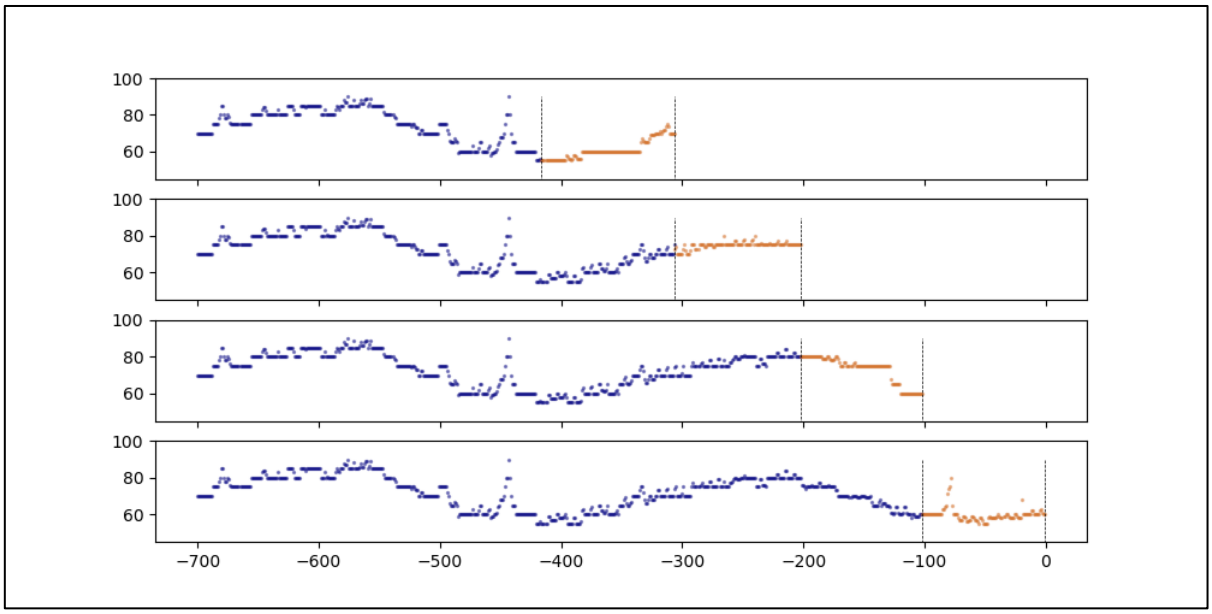
Appendix 2.5. Evaluating booking model base algorithms performance



Appendix 3. Hyperparameter-tuning through forward-chained validation

Appendix 3.1. Pricing model validation MAE per fold

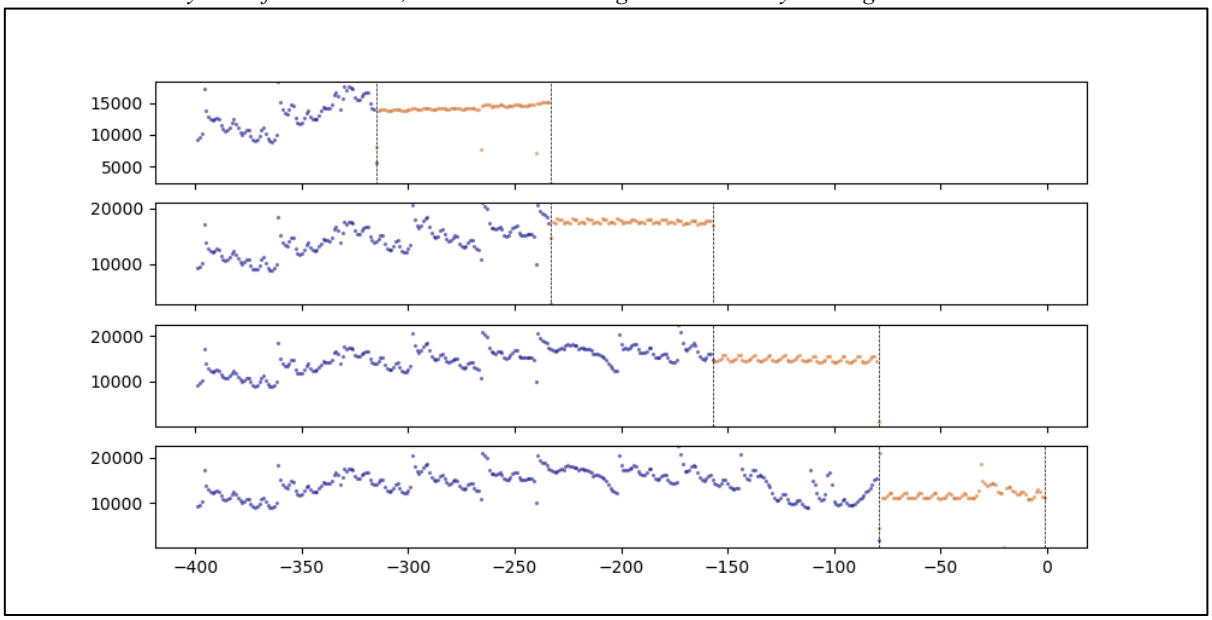
Note: X-axis are days until first lockdown, Y-axis is median price as estimated by winning model. Since counts per day are not equally distributed, it appears like some fold are longer (time) than others, but all folds have the same number of instances



Algorithm	Hyperparameters	F1	F2	F3	F4	Score
LGBMRegressor	{'learning_rate': 0.1, 'reg_alpha': 0}	31.169	31.553	29.630	29.408	30.197
LGBMRegressor	{'learning_rate': 0.1, 'reg_alpha': 0.01}	31.830	31.553	29.630	29.404	30.196
LGBMRegressor	{'learning_rate': 0.1, 'reg_alpha': 0.1}	31.845	31.553	29.630	29.407	30.197
LGBMRegressor	{'learning_rate': 0.01, 'reg_alpha': 0}	44.067	42.397	41.110	44.127	42.545
LGBMRegressor	{'learning_rate': 0.01, 'reg_alpha': 0.01}	44.067	42.397	41.110	44.127	42.545
LGBMRegressor	{'learning_rate': 0.01, 'reg_alpha': 0.1}	44.067	42.397	41.110	44.127	42.545
LGBMRegressor	{'learning_rate': 0.001, 'reg_alpha': 0}	55.455	53.147	52.974	55.738	53.953
LGBMRegressor	{'learning_rate': 0.001, 'reg_alpha': 0.01}	55.455	53.147	52.974	55.738	53.953
LGBMRegressor	{'learning_rate': 0.001, 'reg_alpha': 0.1}	55.455	53.147	52.974	55.738	53.953
DecisionTreeRegressor	{}	19.727	22.578	25.655	22.934	23.722
LassoCV	{'eps': 0.01}	57.549	55.375	55.476	58.146	56.332

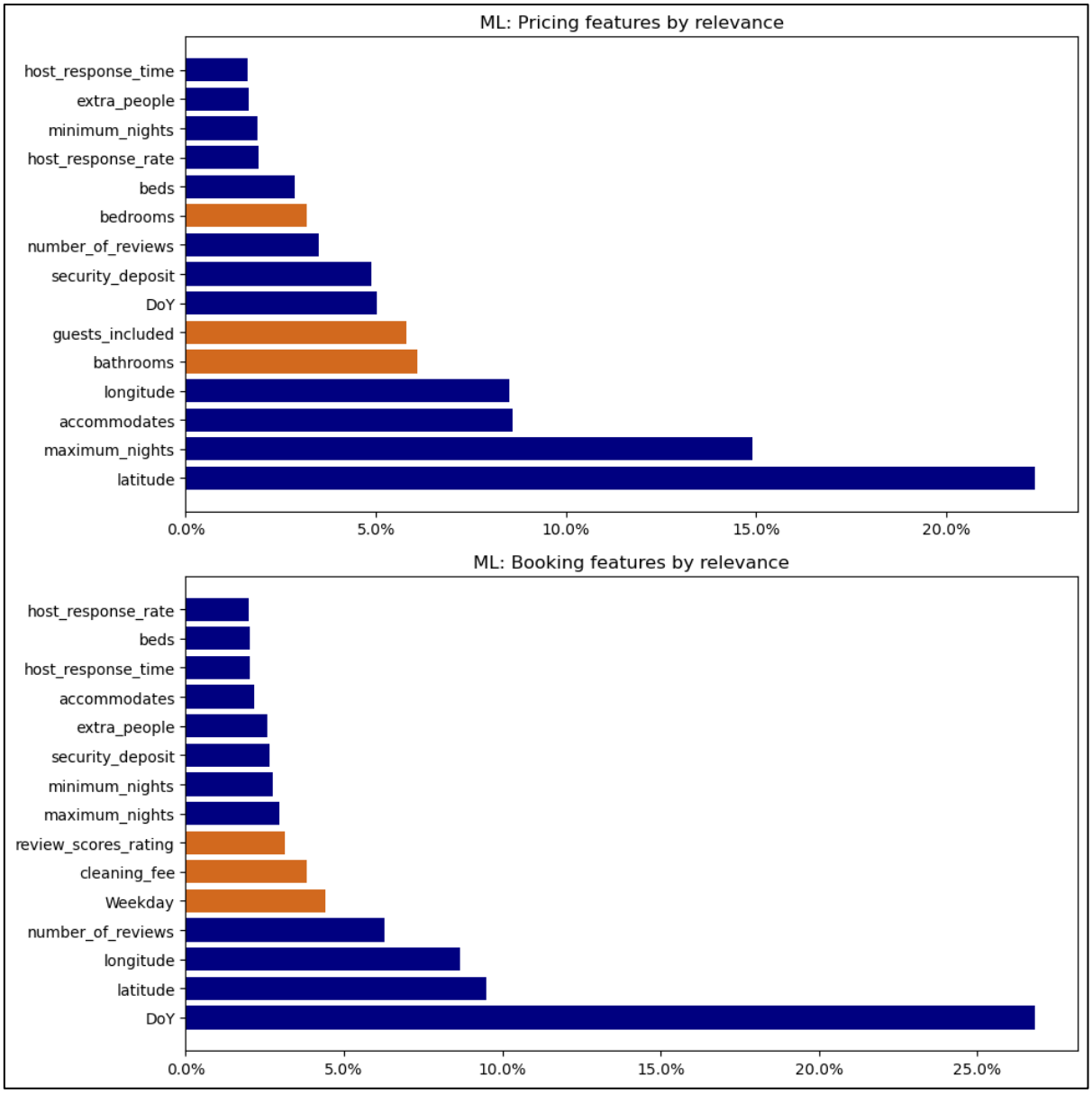
Appendix 3.2. Booking model validation F1-score per fold

Note: X-axis are days until first lockdown, Y-axis is total bookings as estimated by winning model



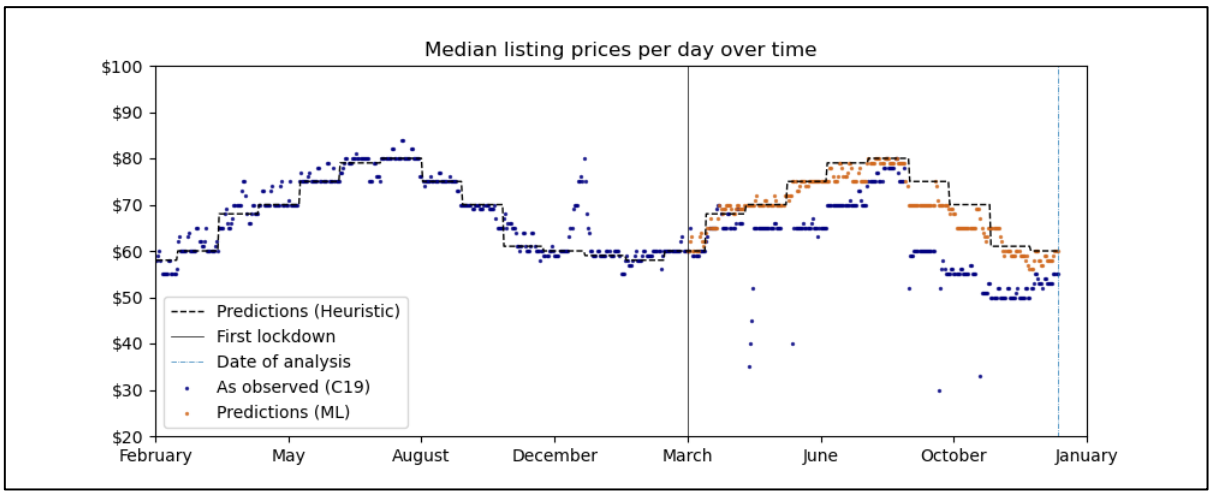
Algorithm	Hyperparameters	F1	F2	F3	F4	Score
DecisionTreeClassifier	{}	0.667	0.702	0.619	0.611	0.644
LGBMClassifier	{'learning_rate': 0.1, 'reg_alpha': 0}	0.305	0.061	0.323	0.607	0.330
LGBMClassifier	{'learning_rate': 0.1, 'reg_alpha': 0.01}	0.306	0.061	0.326	0.606	0.331
LGBMClassifier	{'learning_rate': 0.1, 'reg_alpha': 0.1}	0.305	0.046	0.277	0.606	0.309
LGBMClassifier	{'learning_rate': 0.01, 'reg_alpha': 0}	0.088	0.688	0.691	0.519	0.633
LGBMClassifier	{'learning_rate': 0.01, 'reg_alpha': 0.01}	0.088	0.688	0.691	0.519	0.633
LGBMClassifier	{'learning_rate': 0.01, 'reg_alpha': 0.1}	0.088	0.693	0.701	0.519	0.638
LGBMClassifier	{'learning_ rate': 0.001, 'reg_ alpha': 0}	0.000	0.000	0.000	0.000	0.000
LGBMClassifier	{'learning_rate': 0.001, 'reg_alpha': 0.01}	0.000	0.000	0.000	0.000	0.000
LGBMClassifier	{'learning_rate': 0.001, 'reg_alpha': 0.1}	0.000	0.000	0.000	0.000	0.000

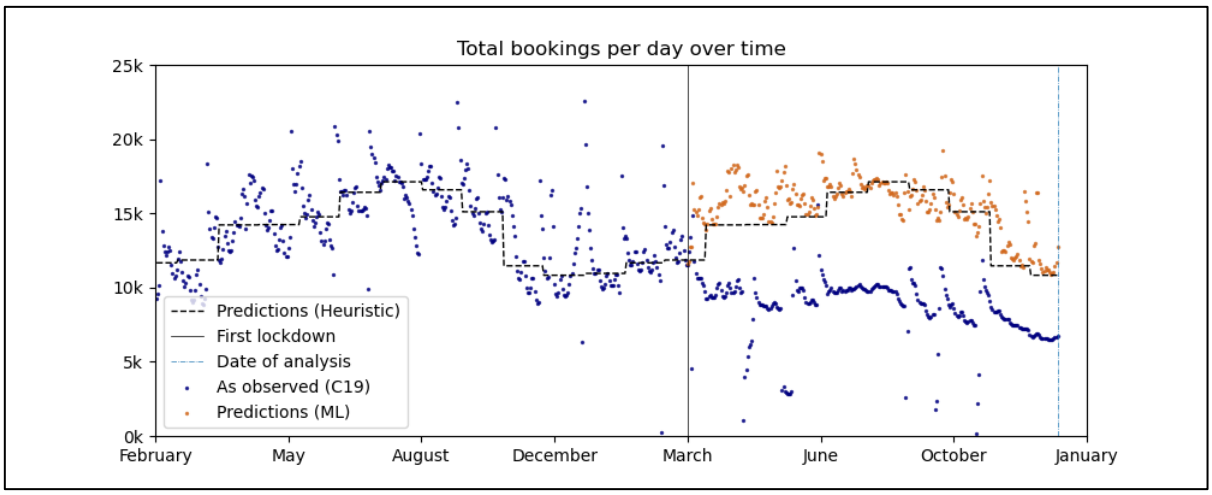
Appendix 4. Feature importance

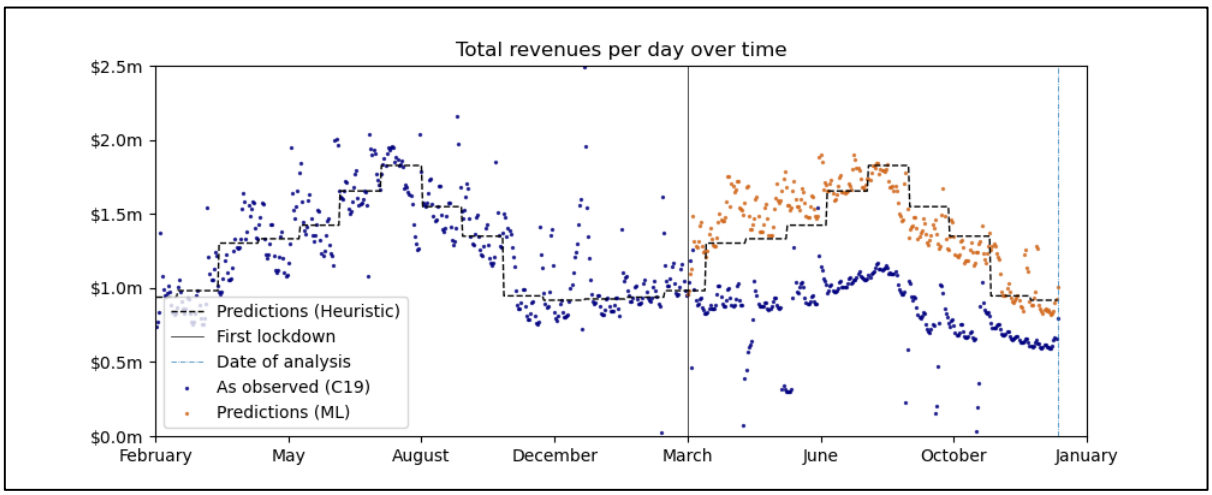


Appendix 5. Predictions over time

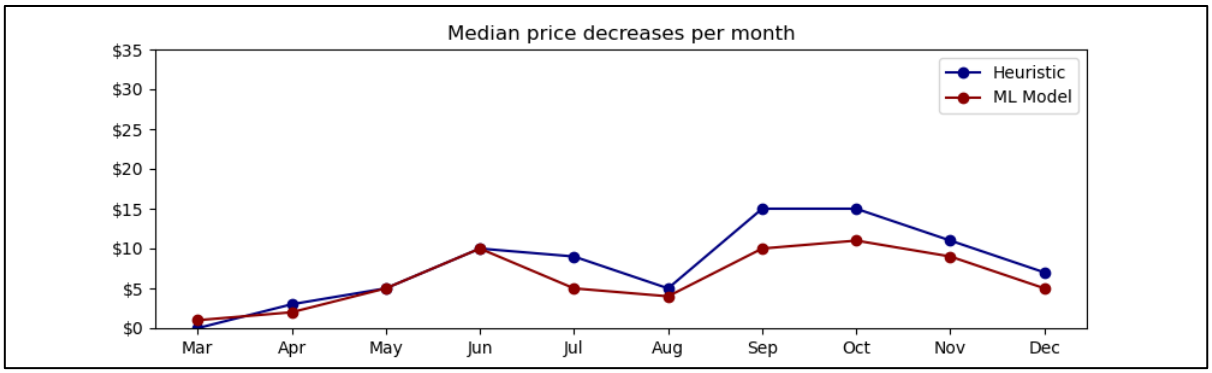
Appendix 5.1. Daily comparison



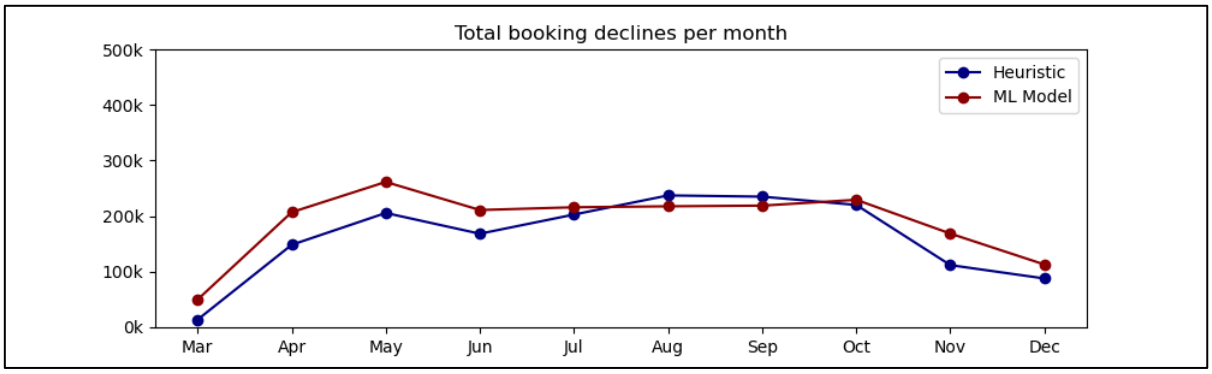




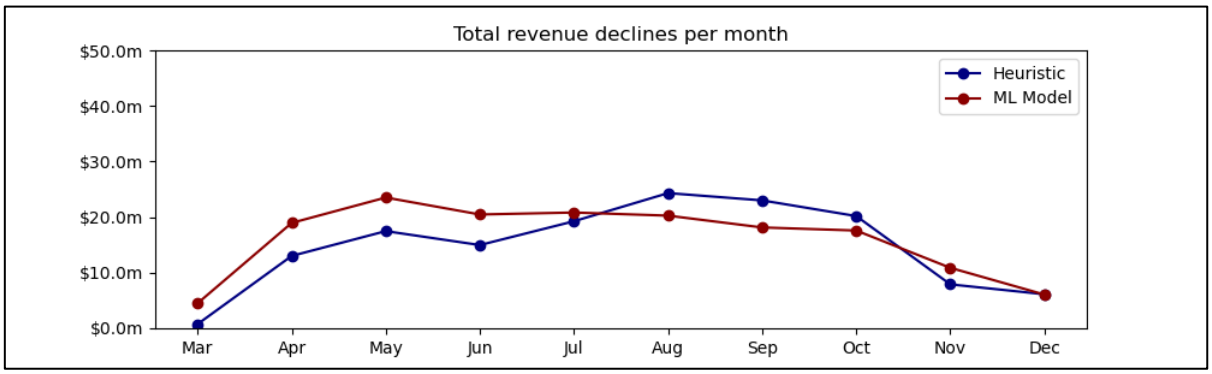
Appendix 5.2. Monthly aggregations: Price



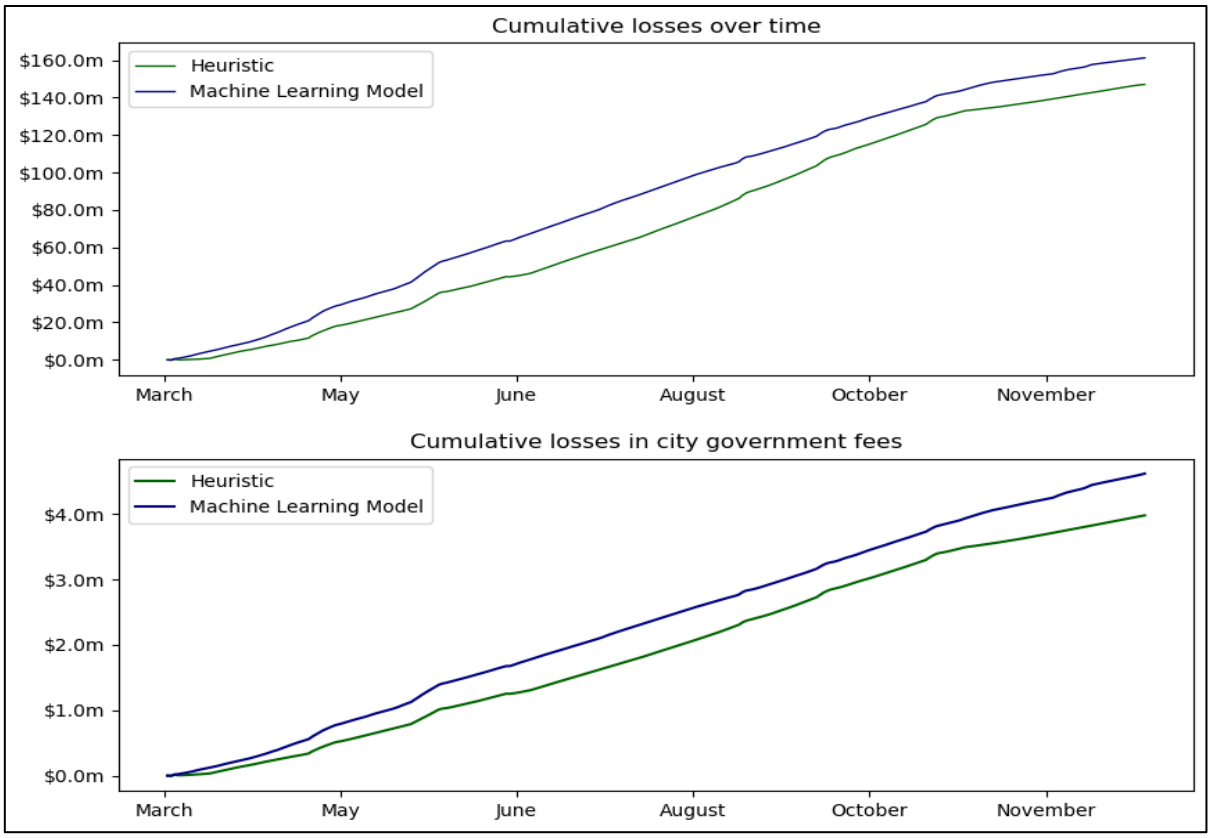
Appendix 5.3. Monthly aggregations: Bookings



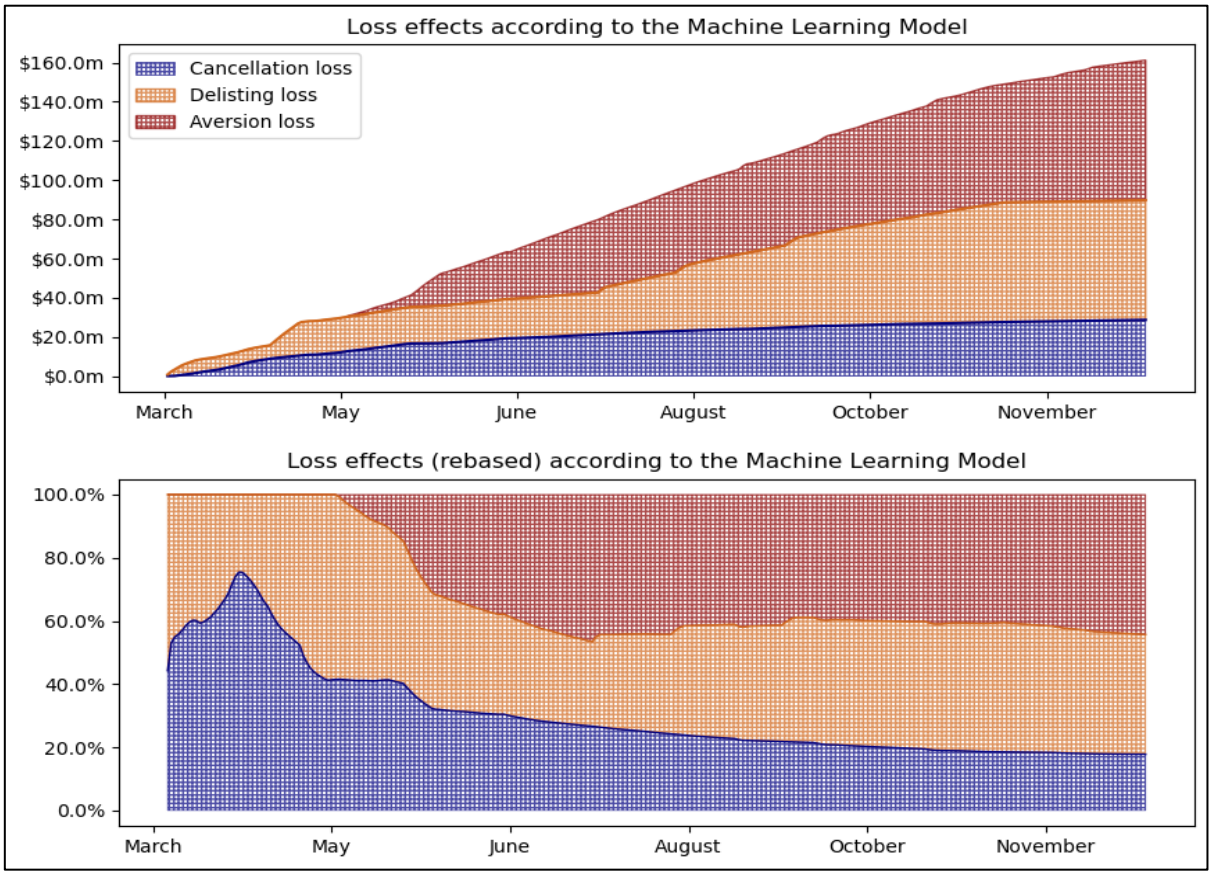
Appendix 5.4. Monthly aggregations: Revenues



Appendix 5.5. Cumulative loss to hosts and city governments

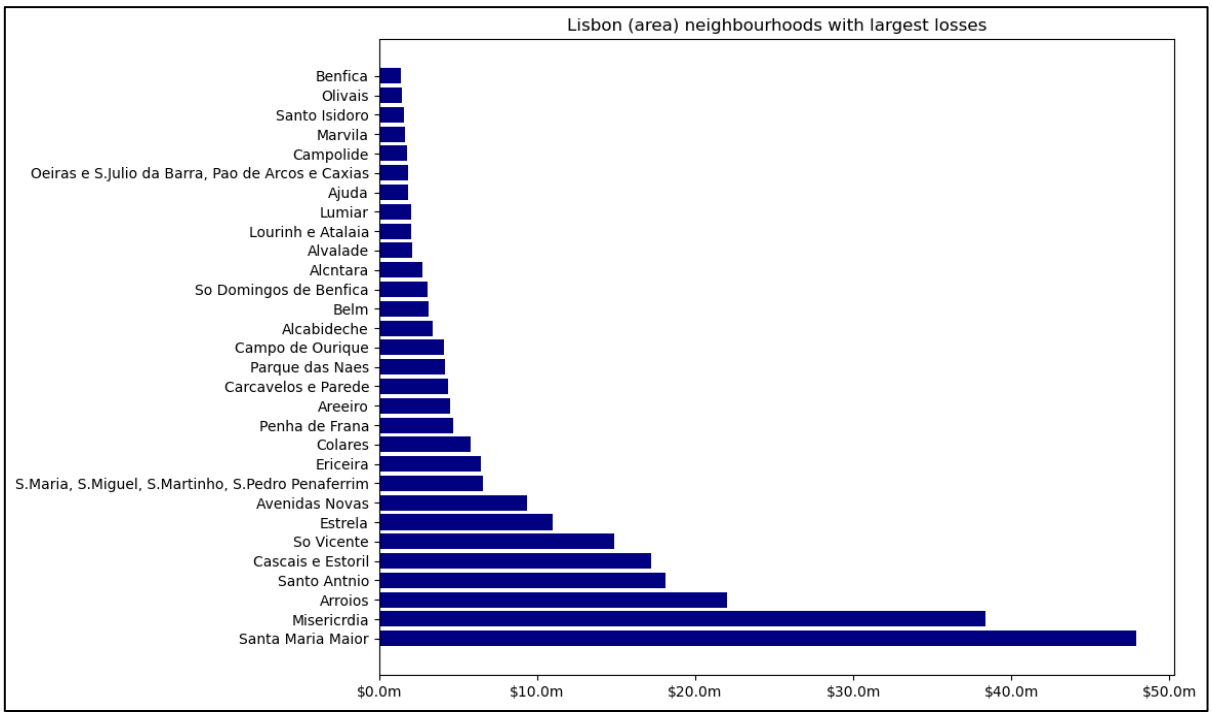


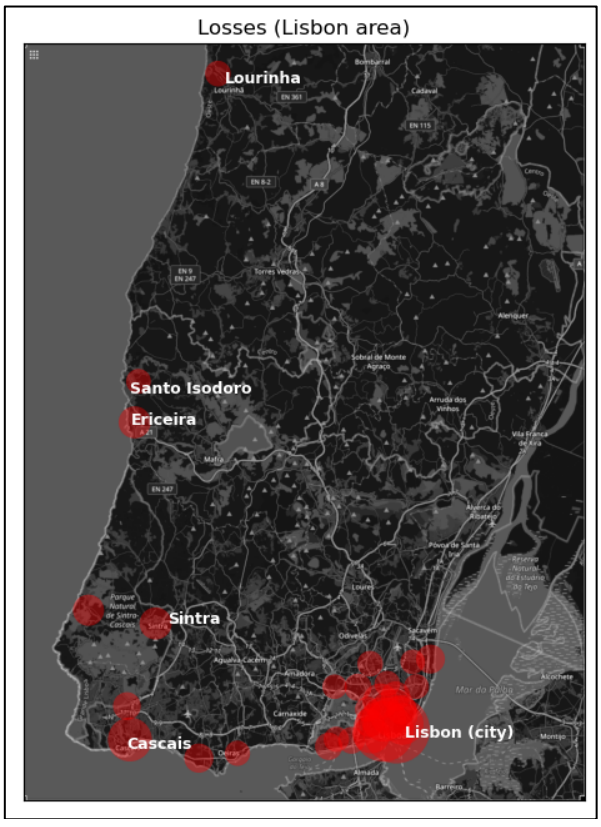
Appendix 6. Impact by forces

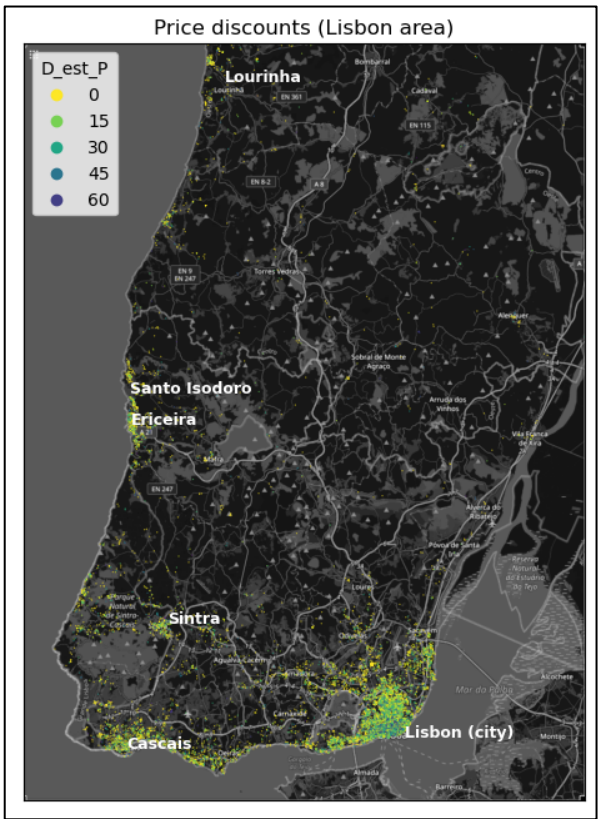


Appendix 7. Impact by location

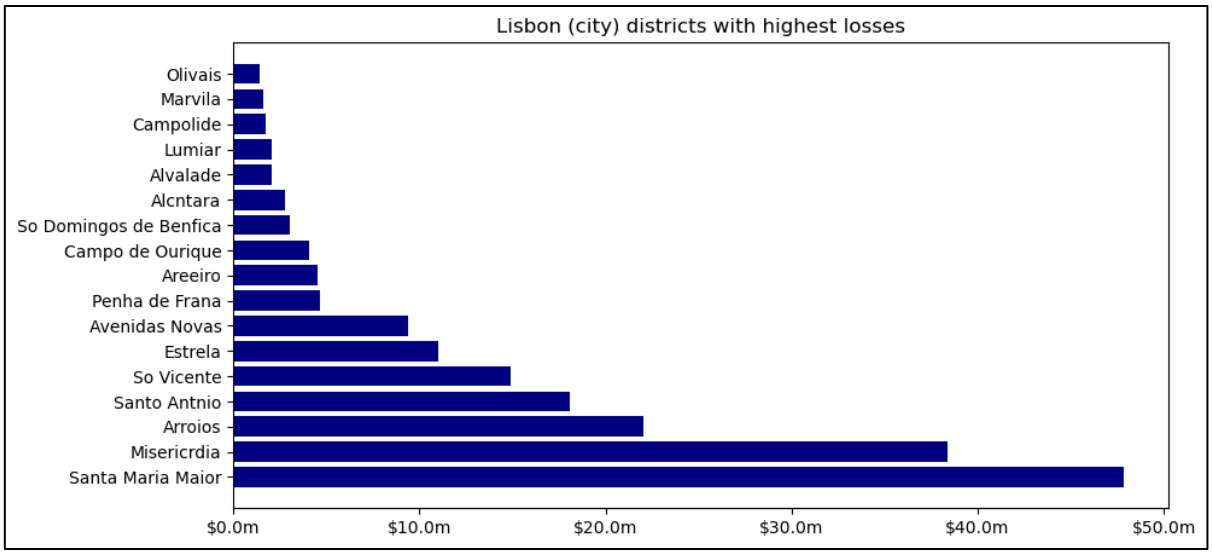
Appendix 7.1. C19 impact on Lisbon (area)

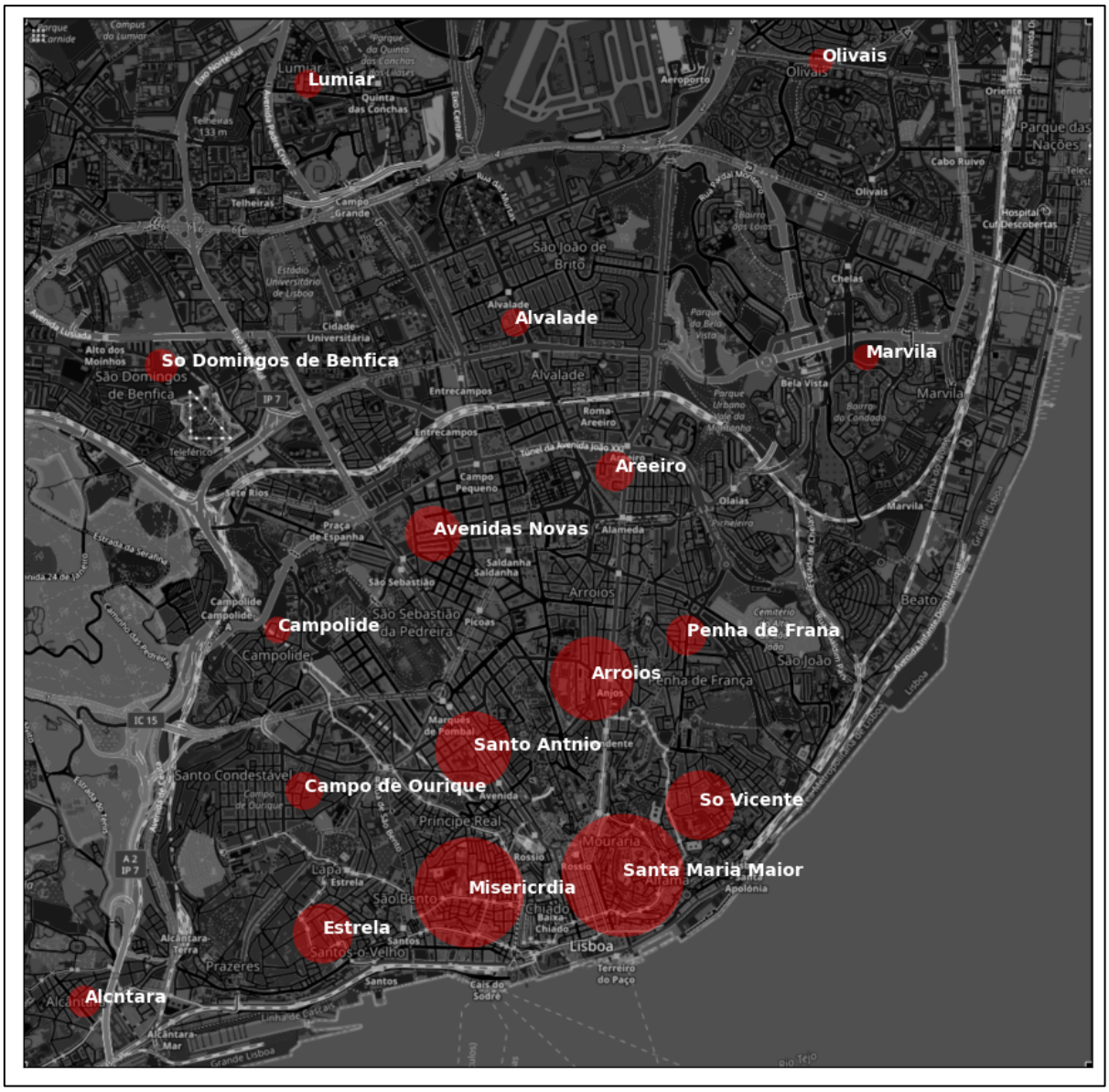




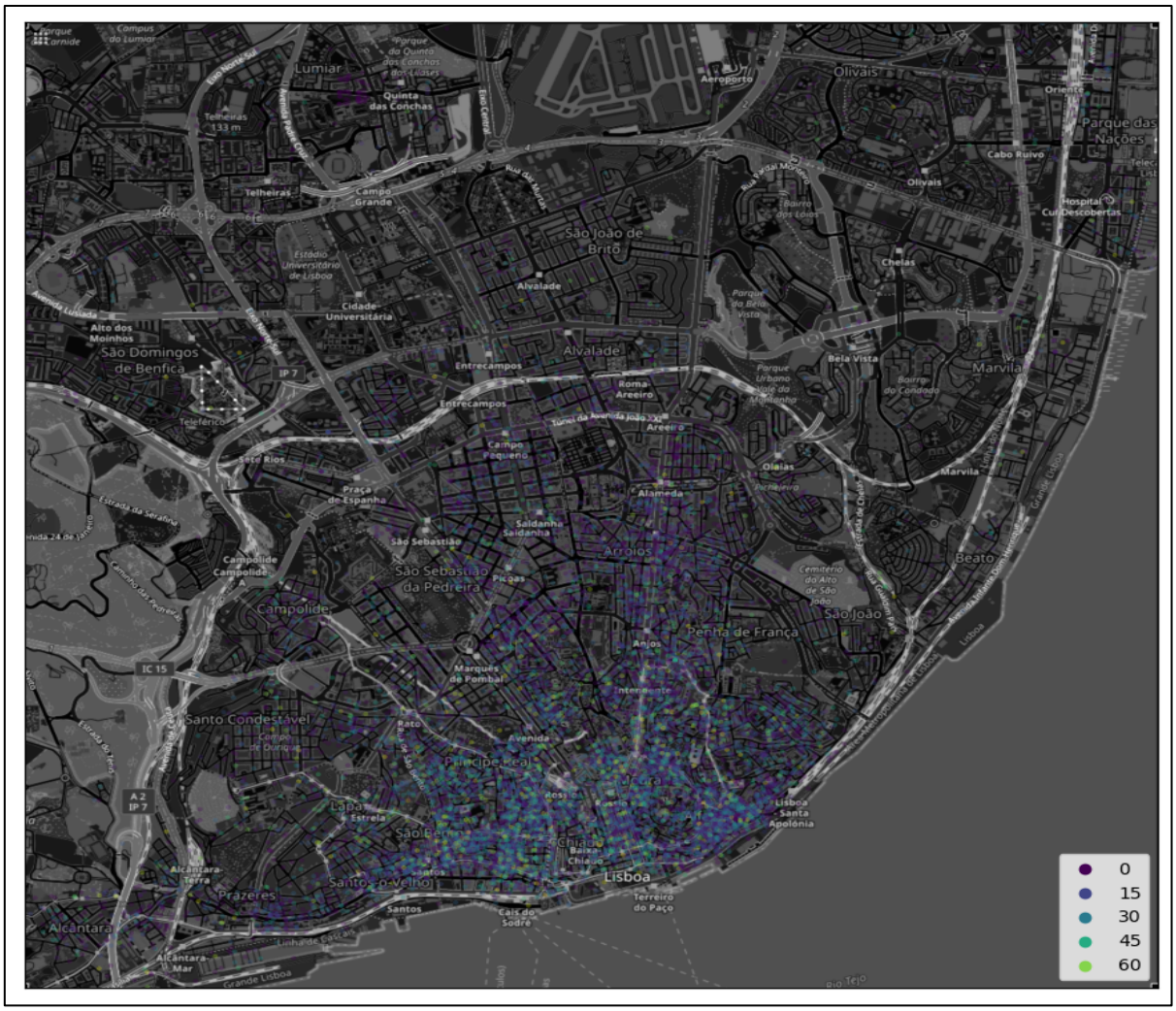


Appendix 7.2. C19 impact on Lisbon (central)

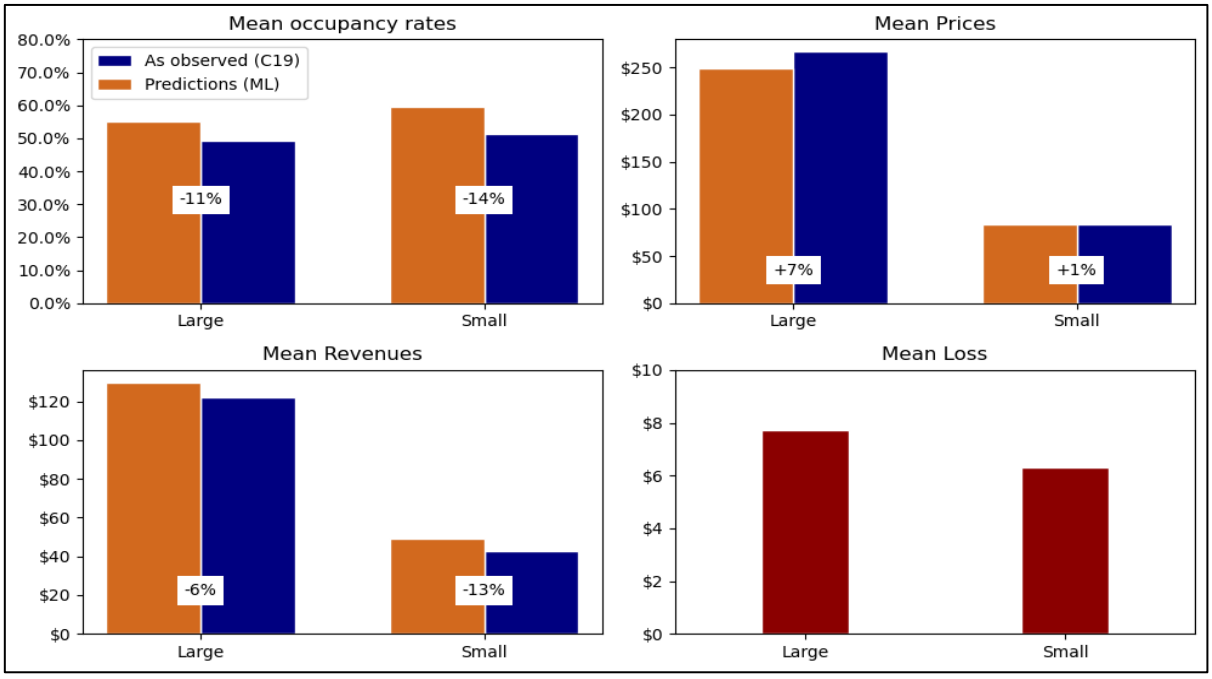




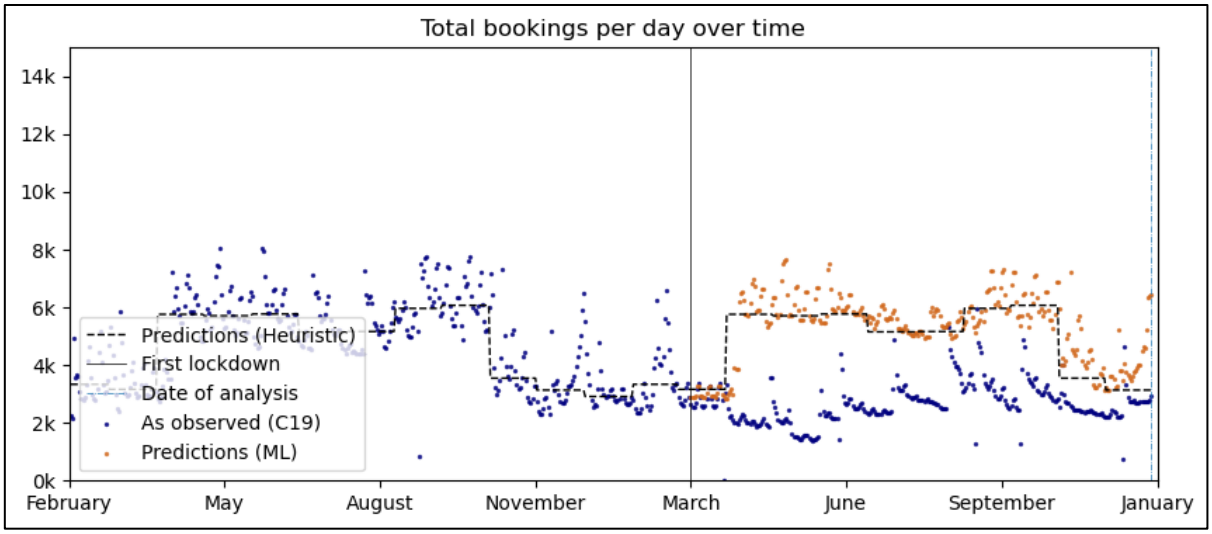
Appendix 7.3. Pricing discounts ($) in Lisbon (central)

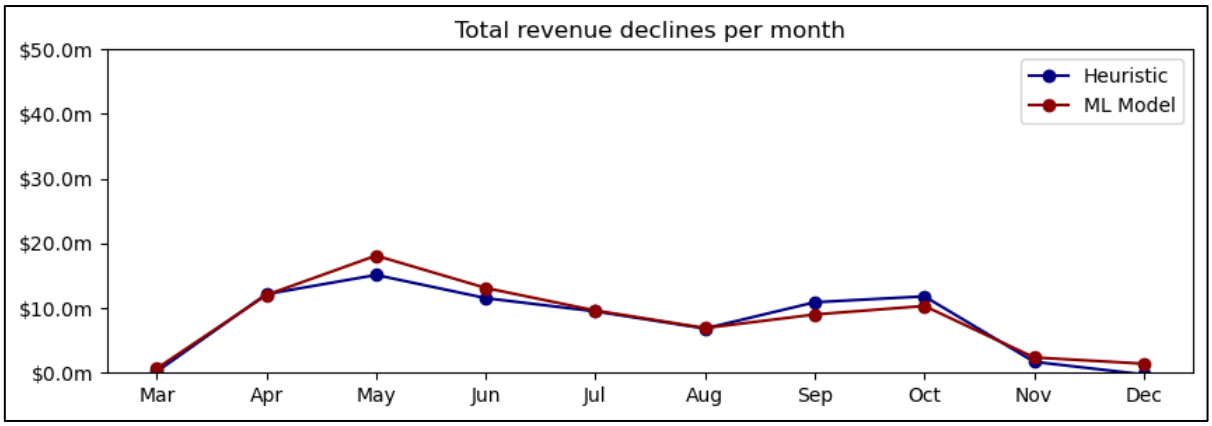


Appendix 8. Impact by size



Appendix 9. Applying the script on data from Venice, Italy





Appendix 10. Python packages used

Use	Packages
Basics	os, sys, time, warnings, datetime, itertools, pickle, operator, collections
Computation	numpy, random, pandas, scipy, math
Scraping	BeautifulSoup, urllib, wget
Plots	cv2, seaborn, matplotlib
Machine learning	Sklearn, xgboost, lightgbmDecisionTreeClassifier

Appendix 11. Table of abbreviations

Abbreviation	Explanation
C19	Covid-19

Abbreviation	Explanation
CPU	Central Processing Unit
GPU	Graphics Processing Unit
HP	Hyperparameters
MAE	Mean absolute error