NOVA IMS Information Management School

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

Operational Impact of Driver Booking at Ifood

Driver Booking Systems in Food Delivery Platforms: data-centered approaches to exploring effects in supply predictability

Tiago José Isidoro Ramos

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

2022

Title:

Subtitle:

Tiago José Isidoro Ramos

MDSAA

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

OPERATIONAL IMPACT OF DRIVER BOOKING AT IFOOD

by

Tiago Ramos

Internship report presented as partial requirement for obtaining the Master’s degree in Advanced Analytics, with a Specialization in Data Science / Business Analytics

Supervisor: Roberto Henriques

September 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Tiago Ramos

Lisbon, November, 2022

ACKNOWLEDGEMENTS

I would like to acknowledge the people that contributed directly and indirectly to the writing of this thesis, without whom I would have not reached this stage.

My academic supervisor, Roberto Henriques, whose passion for machine learning and teaching inspired me to go above and beyond.

The Data team at iFood and in particular Gustavo Pilatti, Mariana Pereira, Marcelo Januzzi, Maxwell Marcos, and Pedro Farias, for being outstanding mentors and teammates. For their dedication, enthusiasm, and support during this process.

The members of the business team; Matheus Zerbetto, Eduardo Mendes, Jessica Naomi, Ana Guimarães, and Isabella Nucci. For their perseverance and commitment to innovation, whose understanding and receptiveness were crucial in accelerating the growth and development of the project.

To the sponsors and witnesses of the project, Erick Farias, and Lucas Nery.

ABSTRACT

Food Delivery Digital Platforms have been investigating novel approaches to leverage drivers' interactions with the platforms to increase operational efficiency. At iFood, one such approach is the introduction of a Booking system, wherein drivers can get increased order priority for scheduling shifts in advance. The theoretical assumption is that this system can increase visibility of drivers' intentions as well as influence their choice of working regimes, leading to an increase in supply predictability. The present internship report details the first steps towards the discovery and validation of the aforementioned assumption, by applying statistical inference tests to the relevant data, as well as training and testing predictive modeling that leverages the new information available by the system and comparing it to current operational models used internally. The results show the newly introduced variables are important, albeit the quantifiable impact is comparatively small. However, the applied models achieve better performance and test scores than the current internal models. Based on the findings of the project and observational insights from previous studies, future steps are proposed, which include the refinement of current operational models, and business initiatives with a positive impact on the importance and quality of specific variables related to the booking system.

KEYWORDS

Food Delivery Platforms; Gig Economy; Driver Shift Booking

INDEX

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

Term	Description
Booking System	The opt-in system implemented by iFood, subject of this thesis. Its outline can be found in the introductory section.
Booking Interface	The in-app interface used through which drivers book specific shifts.
Booked and non-booked metrics (or variables)	Refers to computation of operational metrics (such as deliveries) for the specific group of the driver population (e.g., Booked Deliveries counts the number of deliveries made by drivers who booked that shift).
BAT	Stands for Batch; represents the number of orders batched in a delivery / route.
DLY	Stands for Delay, describes the percentage of orders delivered with a delay (delivery time exceeded estimated or promised delivery time).
DTXX (30,60)	Shorthand-name for Delivery Time; features describing the percentage of deliveries completed under 30 or 60 minutes, respectively.
EDT	Stands for Estimated Delivery Time.
LO	Stands for Logistical Operator; the designation of a locally based company which provides a contract-based workforce to iFood.
Logistical Region	An internally defined geographical delimitation of the company's operation, typically similar to the corresponding geographical region. Currently, Ifood operates on upwards of 500 logistical regions of different dimensions and scope.
Modal / Modality	Describes the type of transportation the driver is using, e.g., bike, motorcycle, car.
OCC	Stands for Occupation; an Operational Metric describing the ratio of the driver's worked and supplied hours. The higher it is, the higher the percentage of logged time that the driver spent *en route*.
RDT	Stands for Real Delivery Time.
Route	Defined as a route proposal to a driver, which includes at least a pickup at a restaurant and a delivery at the customer's locations. A route is considered completed only when both these stages are achieved. A route may also contain more than one order, with multiple pickup and drop (delivery) locations (see BAT).
Segmentation	Internal classification based on the driver's performance and quality indicators. All things held constant; a higher segmented driver has more priority when a route is assigned.
Shift	A named period of time, typically ranging from 2 to 3 hours, representing a specific timeframe of the day. E.g., morning, lunch, afternoon.
Supply Hour	Time that a driver spends logged in the platform, in hours. Operational metric that measures supply.
UTR	Utilization Rate, an operational metric of drivers, measured by the number of delivered orders divided by the provided supply hours.
Worked / Working Hour	Logged time in which a driver is working, i.e., the period between accepting an order and delivering it to the customer.

1. INTRODUCTION

1.1. Presentation

iFood is a Brazil-based company that focuses on online food ordering and delivery, through its online and mobile platform. The platform connects restaurants to delivery drivers and consumers, from whom it receives orders that it communicates to restaurants, optionally providing a logistics service (the delivery of the order) by tasking an available driver with the respective delivery.

iFood does not have a contractually employed workforce. Similar to other gig-based platforms such as Uber, its workforce consists of independent workers who are compensated on a per-gig basis and are available only when logged in the platform. This unique working paradigm fits with the evolving concept of platform and gig economy (as put forth by Malik et al., 2021).

1.2. Motivation for Project

The matching of supply to demand in gig economy-based platforms is a widely researched challenge (Zhong et al., 2019). Despite being a third party to the two actors of supply and demand (drivers and customers), platforms exert influence over both, with added operational and societal factors that can be optimized (e.g., time to delivery, user satisfaction).

From the supply side, platforms are challenged to have driver availability to cover the predicted demand, with added constraints on operational indicators such as delivery times. On the other hand, they must also manage periods of oversupply, when demand is comparatively scarce. This poses a risk to drivers' satisfaction and retention.

To address these supply challenges, iFood has implemented several mechanisms, such as financial incentives (i.e., paying more to drivers in moments of undersupply), and contracts with logistical operators, which guarantee supply hours on pre-defined time slots. Both options represent a significant increase of cost per order.

As such, there has been increasing urgency in exploring alternative systems that are capable of achieving similar results for equivalent or lower costs, while maintaining a stable, satisfying, and predictable driver experience.

The Booking system is one such alternative. By communicating the temporal distribution of predicted demand to drivers, via the availability of scheduling slots, drivers are made aware of supply peaks, as well as time slots with unmet demand that they can take advantage of. Additionally, as drivers must schedule slots in advance, drivers' intentions and their working schedules become more visible.

As of March 2022, Booking is currently available in 16 logistical regions in Brazil, as an opt-in system (i.e., drivers can receive orders even if they haven't scheduled the time slot). To incentivize opting in, booked drivers are given higher priority for order attribution.

As the system is still in early development and implementation, the analysis of the hypothesized benefits represents an excellent opportunity for applied academic research.

The rest of this paper is structured as follows: my contributions are outlined and contextualized in the scope of the project. Then, a literature review is conducted on the application of similar systems in the food delivery ecosystem and the respective impacts. The approaches used and applied methodology are reviewed in the following section. Afterward, the results are presented along with a discussion of the most relevant points and their implications. Finally, we summarize the findings and outline limitations and future steps for the work.

1.3. PROJECT OBJECTIVES

As part of the internship at Ifood, I integrated the Booking task force, as a member of the Data Logistics team. My contributions supported the Business and Product teams' decision-making in terms of future priorities and long-term vision. For the purposes of the required academic work, this thesis will be focused on the projects related to the following research question:

Does the Booking system increase the predictability of Supply at iFood?

This work contributes to the growing body of literature regarding control regimes in gig economy labor platforms (Gandini, 2018), specifically in what concerns numerical and statistical analysis of the impacts of these systems on the supply of these platforms. Additionally, it paves the way for the development of more robust predictive models by investigating the integration of new features.

1.4. CONTEXTUALIZATION

1.4.1. Outline of Operational Model - Drivers

To properly introduce the booking system in the context of the platform, it is relevant to outline the standard operational model for the driver. When signing up for the platform, drivers undergo a screening process (known as onboarding), after which they are allowed to log in to the platform. When logged in and available, they can then accept or reject proposals of routes. Routes are created and assigned based on a complex algorithm that optimizes routes (for example, by batching two orders with very close delivery locations) and selects drivers for assignment based on several factors, including distance to the restaurant (i.e., the pickup location), their segmentation, modality, and others.

Given this context, we now explore the Booking pipeline.

1.4.2. Outline of Booking System

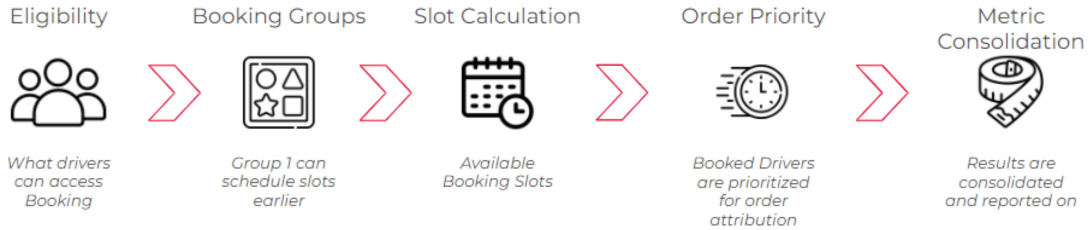

Figure 1 - Depiction of the Booking Pipeline. The booking system follows a sequential cycle repeated weekly. First, eligible drivers are identified; then, they are divided into groups that can book a pre-calculated number of slots. When logging in

during their booked timeslot, the system assigns higher priority to these drivers; finally, metrics are collected, consolidated, and analyzed.

In order, the key moments of the system, highlighted in figure 1, are described as:

1. Eligibility Process: via database querying, a list of eligible drivers is produced. Eligibility for the system is based on business rules such as modality, and logistical region.
2. Division of Booking Groups: The eligible drivers are divided into two groups, based on performance and attendance-based criteria. These two groups are differentiated on how early they are allowed to book future time slots.
3. Slot Calculation: In parallel, a team uses data on predicted demand and other operational metrics to calculate the number of available slots for each shift.
4. Booking: Through the app, drivers can book shifts a few days in advance, subject to slot availability.
5. Prioritization: During the corresponding shift, booked drivers are given additional priority when orders are assigned.
6. Consolidation: An ETL pipeline produces monitoring metrics.

Since it is in rapid development, it is subject to frequent changes. As an example, at the time of writing, the list of eligible drivers is updated weekly (expected to change to a daily frequency in the future). Slot availability is also calculated for the entirety of the following week; slots can currently be scheduled 3 to 5 days in advance (depending on the booking group).

2. LITERATURE REVIEW

This section outlines previous research work carried out in food delivery companies regarding distribution of supply, specifically those that have adopted similar Booking-type systems (wherein drivers can schedule one or more shifts in-app).

Literature on predicting supply (i.e., driver availability) for on-demand app services is quite limited. No thorough academic research on the quantifiable impact on supply predictability of these systems is published externally.

Despite this, publications exist that outline the heuristics behind these systems in companies that have adopted them. The domain knowledge extracted from these experiences is useful to determine basic guidelines for both methodology and interpretation of results.

A very common concept in the literature regarding this subject is that of Control vs Autonomy. It is understandably prevalent in the topic of gig economy. The dichotomy represents the balance between gig workers' agency to choose working hours and the needs of the company using those working hours. When companies exercise greater control (for example, by restricting working hours, imposing penalties, or using financial incentives to increase attendance), it comes at the cost of the worker's perceived autonomy, and dissatisfaction (Stewart et al., 2020).

2.1. SHIFT BOOKING IN OTHER COMPANIES

In their research, Ivanova et al. (2018) reference shift booking systems as a controlling mechanism of driver supply in two food delivery companies in Germany: Foodora and Deliveroo.

Foodora employs a strict Booking-only system, where drivers can only work in pre-scheduled time slots. In both companies, drivers are assigned priority groups to schedule shifts, based on their performance. This leads to a self-reinforcing mechanism wherein the lowest-prioritized drivers are unable to work at the only available times and therefore are unable to raise their priority, as the companies use mainly attendance-based metrics to sort these priority groups. Moreover, in some cases, drivers had shifts automatically assigned. This system was perceived as punitive and unfair by drivers, leading to a reported increase in driver churn. In 2016, Deliveroo invested heavily in recruiting new drivers, with a focus on guaranteeing sufficient supply for all shifts.

In a research paper that reported similar findings, Heiland (2021) explores the same two companies.

At Deliveroo, some capacity of the system to influence supply elasticity was observed, via a booking-only working regime and penalties on autonomy: "(...) some riders accepted every offered shift, regardless of its profitability or attractiveness, to improve their status and to become able to choose shifts as part of the first group (…)".

At both companies, interviewed drivers declared a loss of flexibility and volatility in income, imposed by the company's operational control of shift allocation and restrictions.

The paper also highlights a connection between worsened working conditions (when control is high and autonomy is low) and increased driver churn for companies with these types of systems.

Moreover, it finds drivers do not work in less-preferred time slots because they provide larger earnings, but rather because they feel forced to increase their stats and priority to book future shifts. It also highlights how “labor markets are location-specific and as such they influence the effects of control regimes;”, and “...different employment models result in different control needs”. For example, this level of control is facilitated by the large reserve of willing standby drivers. Were it not the case, the high turnover would inevitably lead to a shortage of drivers. Additionally, it is argued that time slot allocation (described as working hours regimes) is effective at controlling the supply of the examined food delivery companies.

Drahokoupil & Piasna (2019) analyzed Deliveroo’s operation in Belgium. The case study is relevant as it highlights the existence of a booking system under a particular set of labor laws that benefited drivers’ working conditions (with fixed hourly wages). The study found most of the labor force was composed of students looking for additional income, attracted by the flexibility of working hours. In surveys, the choice of working shifts was mainly motivated by the driver’s schedule, rather than payout maximization.

In Australia, Deliveroo employed similar strategies and systems (Veen et al., 2019). The research identified manifestation of drivers’ agency through resisting behaviors, such as reworking the technological systems, ignoring attempts to direct supply, or evading shifts. It additionally highlights the importance of the geo-social context in the configurations of these systems but does not present the potential impact of the system on the predictability of workers’ hours.

In a qualitative study conducted between 2018 and 2019, Griesbach et al. found indications of decreased satisfaction, perceived autonomy, and quality of working conditions for Instacart, when compared to other food delivery companies in the same region. These differences were mainly attributed to Instacart’s usage of an exclusive shift-booking regime, requiring drivers to book time slots a week in advance to receive orders.

In the aforementioned cases, direct influence of driver’s supply hours was found only in companies that employed strict control regimes (such as allowing drivers to work only in booked slots). Impacts on control were less clear in companies that employed hybrid schemes (shifts with booked and non-booked drivers) such as iFood. Despite this, the literature points towards supply being relatively inelastic, i.e., if a driver is forced to choose between an unwanted shift (even with potentially better payouts), and not working, most drivers will choose to not work, and eventually, leave the platform.

3. METHODOLOGY

3.1. Framework

Considering the recency and quantity of available data, as well as the lack of foundational research on the subject, the choice of methodology was driven by an objective of discovery. In this sense, we opted for an investigation based on simpler models, techniques, and heuristics, supported by strong domain knowledge of the industry. This approach allowed for targeted exploration of the baseline assumptions of the system.

A CRISP-DM (CRoss Industry Standard Process for Data Mining) based approach was chosen as a framework. CRISP-DM is a widely applied methodology (Schröer et al., 2021), allowing for extensive iteration between its six well-defined stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The steps taken for each of the stages are described in detail in the following sections.

Additionally, following the business understanding stage, two distinct objectives were considered:

1. Investigate the influence of booking variables on supply hour (i.e., extracting information about the causal relationship);
2. Investigate the impact of booking variables on the predictive power of models (predictability).

The applied methodology and modeling is similar for both objectives, with key differences in the feature selection stage and the emphasis placed on the interpretation of feature importances and model errors.

For the causality framework, the feature set represented variables with same-day values. By operational design, and by domain knowledge, the direction of causality between independent variables and the dependent variables is well understood, lending credence to interpretations based on feature importances[1].

For the predictability framework, the feature set was chosen based on internal production models (benchmark models). If the hypothesis that booking variables increase predictive power holds, then we should be able to train models that achieve better scores than the benchmark models, for the same time range. Additionally, the booking-related variables should be relevant and important to the best-performing models.

As such, we compared the model scores to those of the benchmark models. We also extracted the feature importances of the best-performing models.

This type of approach was considered appropriate given the time constraints. Additionally, it provided a good starting point for more statistically robust analyses, enabled by the constructed pipeline and available data, in the future.

[1] Further challenges to causal inference in this project are explored in the limitations section of this work.

3.2. BUSINESS UNDERSTANDING

One of the proposed objectives of the Booking system was to increase the predictability of driver supply. For the business team, this notion has a two-fold meaning. On one hand, it describes the objective of using the system to effectively manage driver's supply hours (analogous to a causal effect). On the other hand, it describes the objective of achieving more accurate forecasts of the driver's supply hours (regardless of causality)[2].

Based on domain knowledge, the team proposed these objectives could be achieved by:

- Direct influence via the definition of available slots; i.e., the cap on available slots leads drivers to schedule for less crowded, or available, time slots;
- The informative power provided by the driver's interactions with the booking tool. For example, the average attendance rate of a logistical region, as well as how many slots are filled, should contain information about drivers' behaviors that were not available before.

A few challenging considerations emerge for these theoretical proposals:

- The first point is a matter of supply elasticity. Even when booked slots are full, the driver can choose to work the same time slot (for example, out of habit, or if they perceive the hourly earnings are advantageous despite having lower prioritization), or just not log in at all. This elasticity is unknown; therefore, the theoretical causal effect of the system is unclear.
- The new variables may not bring new relevant information to models if the captured behavior can be inferred from other, previously available variables.

Due to the recency of the system itself and the lack of available data, little research had yet been done to explore and provide evidence of the impact of the Booking system in supply control and predictability (or lack thereof), or of the above points. Moreover, the specific nature of this ecosystem means that general research on the topic is only tangentially applicable in this context.

3.3. DATA UNDERSTANDING

3.3.1. Data Structure - iFood

At iFood, data undergoes rigorous ETL processes. It is initially collected in a centralized data lake, from which it is extracted through data ingestion pipelines into "raw" databases, where data is stored as is, having gone through LGPD validations (i.e., data protection validation, when personal info is dropped or anonymized). Data from these sources can be used by a restricted part of the user base. Generally, pre-processing of the most important and relevant data is coded through Databricks notebooks and automated via Apache Airflow, generating "curated" and up-to-date databases, which are greenlit for usage within the company.

Additionally, a sandbox area exists, allowing for the creation of temporary databases without any stringent quality criteria, used for data exploration tasks, testing of curation pipelines before

[2] Emphasis was placed on the increase of predictability via the additional information provided by the system, i.e., the success of the booking system was not dependent on proving a causal effect on supply.

deployment, or maintenance of proof-of-concept pipelines that are subject to very frequent changes. This structure is depicted in Figure 2.

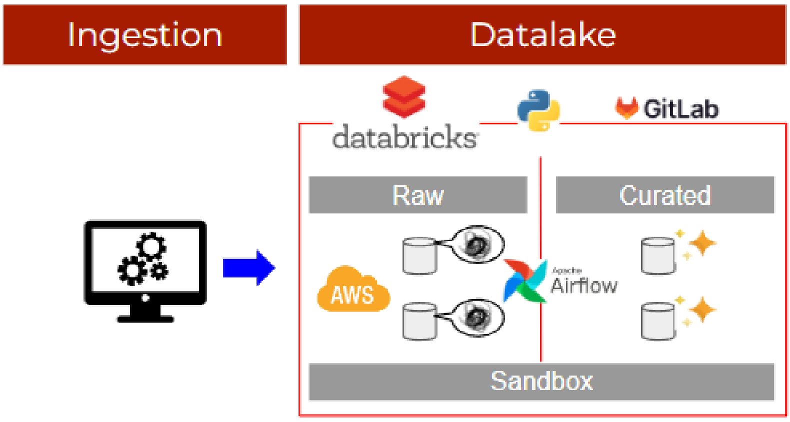

Figure 2 - Depiction of Data Structure Systems at Ifood (simplified). The company utilizes different services of data ingestion. The datalake is divided in three broad areas; a "raw" zone, a "sandbox" zone, and a "curated" zone, with different levels of data quality specifications. While the "sandbox" zone exists independently to hold datasets of a temporary and experimental nature, "curated" datasets are maintained by automated pre-processing pipelines and tasks (via Airflow), sourcing data from the "raw" zone. These pipelines are created and maintained with Amazon Web Services, Databricks, Gitlab, and Airflow.

3.3.2. Benchmark Models

iFood currently employs operational forecasting models to predict driver's supply hours. Considering the purposes of the research, these models represent a suitable benchmark and starting point for the development of new models.

The benchmark models used are Random Forest Regressors from the sklearn package. An individual model exists for each logistic region and shift, trained and tested for data on each combination of those features[3].

The features used in these models include information on supply hours (lagged feature), precipitation intensity and temperature (proxy for likelihood of rain), whether the day is a holiday (or if there was a recent holiday), the day of the week, and driver's registration and activation plan (history of the number of drivers and expectation of new drivers entering the platform). The data used for the construction of these features is sourced both from the company's datalake (ingested), as well as specialist pipelines curated by live operations expert teams.

Previous insights from this internal project indicate the most important feature is the previous week's Supply Hour, with all others being comparatively uninformative.

Configuring the pipeline or tweaking these established models was out of the scope of this project. For this research, we extracted the predictions and corresponding real values for SH for all relevant time ranges, and for the same logistical regions. The error scores were then computed.

[3] The duration and timing of the shifts used in these models differs from those employed in the booking system.

3.3.3. Dataset Description

In parallel, a dataset containing all the necessary variables to train and test all models was constructed. All data were extracted from the company's data lake, as outlined in section 3.3.1.

These data can be broadly categorized as operational data and booking-related data. This is a relevant distinction, because while operational data can be directly extracted from curated databases and pipelines "as is", booking-related data existed solely in raw and sandbox databases, a fact attributable to its recency[4].

The curated databases were sourced to extract operational data with known links to supply hour, including some of the features present in the benchmark models. The constructed operational dataset contains information on demand (orders and deliveries), supply hours, driver occupation, delivery times (and delays), the likelihood of rain, and consumer satisfaction. The dataset contained data from January 2021 to January 2022, a total of 220311 rows (each corresponding to one of 5 daily shifts of a logistical region on a specific day).

For booking-related data, a consolidated dataset of booking key metrics on the granularity of logistical region and shift was built. These data were extracted from sandbox databases, and include information on booking slots (available, filled, attended), attendance rate, permanence rate, and booked share of supply hour. The dataset contained data from November 2021 to January 2022, a total of 3326 rows (each corresponding to a shift of a logistical region on a specific day). A description of the pre-processing steps taken for this dataset is covered in the following section.

The two separate datasets (operational data and booking data) were then merged by logistical region and shift, containing a single unified dataset with all the necessary data to train and test predictive models after the introduction of booking. Due to data inconsistencies, the final dataset contains 3130 rows, from November 2021 to January 2022.

The operational data from January 2021 to November 2021 was kept in a separate dataset, for the training and testing of models for a pre-booking time range.

It is important to highlight the cut-off date for the introduction of Booking in November 2021, as it does not correspond to when the system was introduced in the region, but rather the moment from when data started being collected and recorded. This generates imbalances in the dataset as some logistical regions have decreased representation in the dataset. Consequences of these imbalances are explored in the limitations and future work section.

3.3.4. Exploratory Data Analysis

In parallel, the data was explored and examined, to identify relationships between variables and visualize their respective distributions.

Figure 3 depicts the general correlation matrix for the features of the dataset, after some preliminary feature selection and engineering (like the included booking features).

[4] Work on the improvement of existing pipelines is assigned to a different team and outside the scope of the internship.

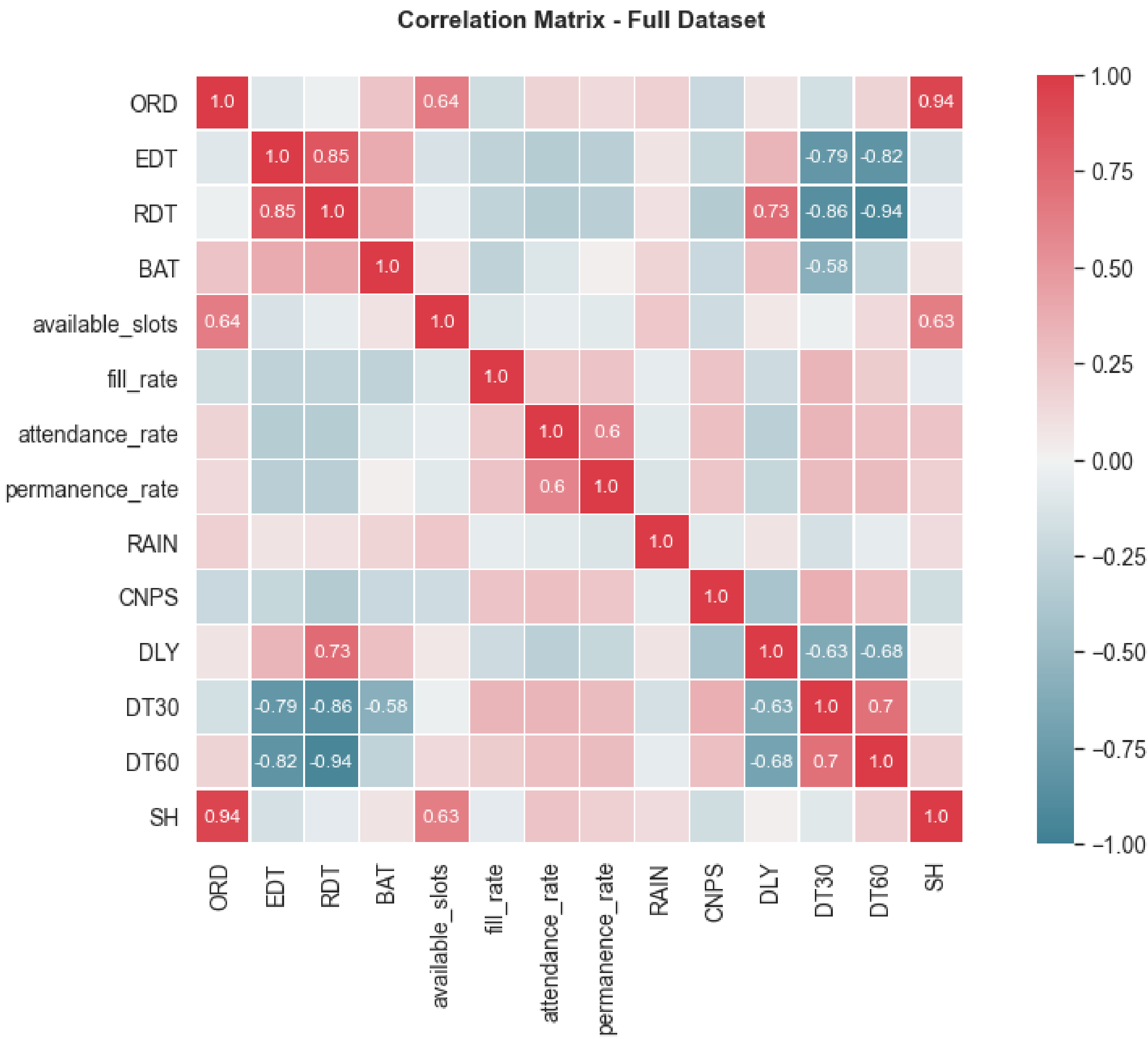

Figure 3 - Correlation Matrix for the Features in the Full Dataset. The heatmap presents values for the highest and lowest correlation values observed. Notably, the variable ORD presents near-perfect correlation (0.94) with the dependent variable. EDT and RDT present high correlation (0.85) and are inversely correlated with both DT30 and DT60. BAT presents inverse correlation with DT30 only. Available_slots presents correlation between both ORD (0.64) and the dependent variable (0.63). The other three booking-related variables (fill_rate, attendance_rate and permanence_rate) present no strong correlation with other variables, except for attendance_rate and permanence_rate, with a correlation of 0.6 between each other. RAIN and CNPS also do not present strong correlation with any variable. DLY is strongly correlated with RDT (0.73), and inversely correlated with both DT30 (-0.63) and DT60 (-0.68).

The insights gleaned from Figure 3 are consistent with domain knowledge regarding the operational models of food delivery. Specifically, the extremely high correlation between ORD and SH are indicative of the overwhelming predictive power of that variable in relation to the dependent variable. This is also visible in the available_slots variable, which is correlated to both ORD and SH, and is calculated using an estimate of ORD.

Figure 4 depicts a histogram of the available data for each logistical region over the considered time.

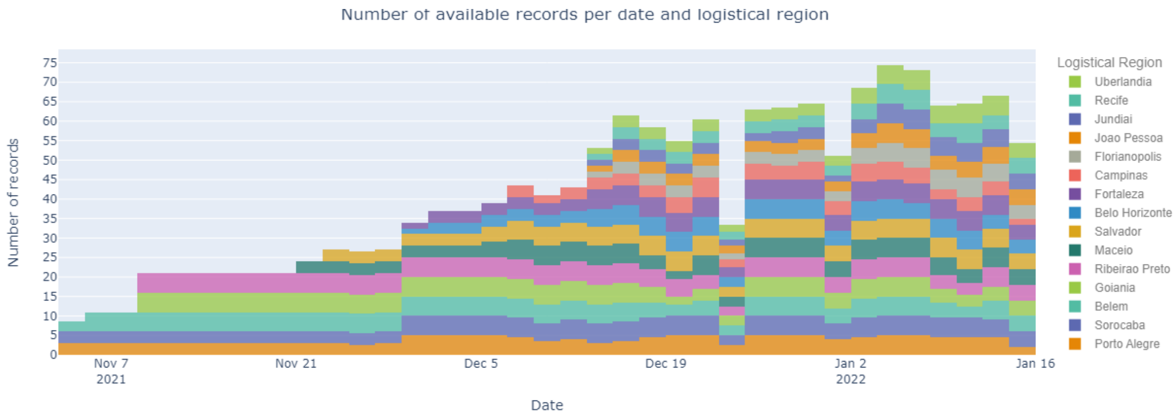

Figure 4 - Histogram of number of records for each region, over time. The histogram depicts the available number of records for each of the 15 logistical regions, for each day, starting November 1st, up to January 16th. The legend order is closely aligned with the respective logistical region, e.g., the bottommost item in the legend corresponds to the bottommost color of the histogram. Each logistical region will have at most 5 records per day (one for each shift). Most logistical regions started with booking on three shifts, gradually increasing the system coverage. This is observable in the datapoint of December 1st. The dips in the histogram represent days for which data was unavailable or dropped from the final dataset.

Significant gaps exist in December and January due to system errors, in addition to the frequent changes in available records due to the gradual introduction of Booking to new logistical regions or new shifts in previous logistical regions. A bar plot of the total number of records for each logistical region can be found in Appendix A.

Figure 5 and 6 depict a line plot tracing the daily 7-day rolling average for the booking metrics, for the five most represented logistical regions in the dataset.

Figure 5 - Evolution of booking metrics (permanence, fill, attendance rates) over time (7-day rolling mean), top 5 logistical regions (in number of datapoints). Dates range from the beginning of November to January 15th. Permanence rates show little variability, with a slight decreasing trend, especially accentuated for the green region. Fill rates are mostly stable for three regions, with sudden spikes and drops for the green and orange region, which follow a similar trend starting on the beginning of December. Attendance rate shows a similar variability to permanence rate (about 0.2), with the blue, red and purple region presenting a very similar evolution over time, and the green region with a significant dip in the beginning of December.

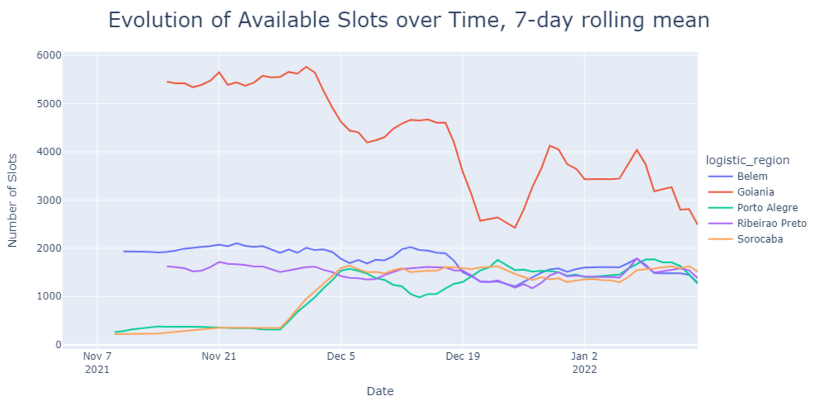

Figure 6 - Evolution of booking metrics (available slots) over time (7-day rolling mean), top 5 logistical regions (in number of datapoints). Dates range from the beginning of November to January 15th. The largest region (red) shows the largest adjustments over time; from about 5500 slots in November and December to just about 3000. The number of slots for the green and orange region was adjusted upwards in December, to similar numbers as the other two regions (purple and blue), ranging from 1000 to 2000 available slots.

Overall, no specific pattern or trend is observable, except for a general decrease in permanence and attendance rate in the beginning of December, followed by a spike around the turn of the year

(January). Additionally, trends are not generally consistent across all logistical regions, albeit the purple and red regions' evolution of permanence, fill, and attendance almost perfectly overlap.

It is also relevant to highlight that the green region shows significant decreases in all rates in the beginning of December, with little recovery. This behavior is partially justified by the increase of available slots, which occurred at the same time (seen in Figure 6). The same increase in available slots occurred for the orange region as well, but the observed effect on the rates was less pronounced.

Although exploring reasons for these patterns is outside the scope of this work, there seems to be some correlation between available slots and rates for some specific regions, which is not captured in the heatmap of Figure 3, perhaps related to the different entering strategies the business team implemented for these regions.

3.4. DATA PREPARATION

3.4.1. Pre-processing

For the operational data, given the quality validation requirements of the databases, pre-processing was necessary only to adjust for discrepancies in time ranges for shifts between the booking system and the company's internal standards.

For the booking-related data, the following pre-processing steps were taken:

- Date types and formats were standardized across tables (including fixing mismatched time zones from different data sources);
- Capitalization and format of logistical regions were standardized;
- After compiling information for booked and attended slots, shifts with no attendance were dropped (7 records).

Afterward, the two datasets were merged, keeping rows where logistical region and shift were available in both datasets. The unified dataset was then checked for consistency by comparing the values between features available in both sources, such as the number of orders for a specific shift (available in sandbox databases for booking, and in the operational curated datasets sourced). Rows with significant discrepancies were dropped (total of 190 records).

One-hot encoding was utilized for the transformation of categorical data.

Data were scaled for linear and k-nearest neighbors models, using Standard Scaling. The same treatment was not applied to the other models (ensemble trees), due to their relative non-sensitivity to variance in the data.[5]

Outliers were detected using Isolation Forests (Tony Liu et al., 2008) (as implemented in the sklearn package). This technique is shown in the paper to be robust in general application and model agnostic, which is useful in the scope of this research. A level of 0.05 was set for contamination, to remove the most extreme cases without significantly reducing the amount of available training data.

[5] Time constraints limited my testing capacity to test different scaling methods; given preliminary results favouring the models using unscaled data, this direction was not followed.

3.4.2. Feature Extraction and Selection

One distinction between the two frameworks considered is in their feature sets. Selection for both frameworks was heavily influenced by domain knowledge (via internally developed operational simulators), with the benchmark models as a starting point.

For the causality framework, the included operational variables form a cohesive picture of the amount of supply time each order takes (BAT, EDT, and RDT). DT30, DT60, and DLY (see glossary) were excluded; theoretically, these features have high correlation with the aforementioned variables, and the information contained can be derived from the others. In practice, their exclusion did not have an impact on any of the model's scores.

For the booking variables of the framework, the features were engineered to represent the same metrics the business team used: available slots, fill rate, attendance rate, and permanence rate. This facilitates the interpretability of the results and translating insights into business decisions (further discussion can be found in the section regarding future work). Additionally, by using rates, the full relationship of the booking variables can be easily expressed as:

*Booking SH = Available Slots * Fill Rate * Attendance Rate * Permanence Rate * SH per slot*

The features and corresponding descriptions for this framework are outlined in Table 1.

Feature	Description
weekday	Categorical feature representing the day of the week. (Monday-Sunday)
logistic_region	Categorical feature representing the logistical region of the observation.
shift	Categorical feature representing the shift of the data point (morning, lunch, afternoon, dinner, night)
ORD	Numerical feature representing demand, i.e., the number of orders.
EDT	Numerical feature representing the expected delivery time of an order, in minutes.
RDT	Numerical feature representing the real delivery time of an order, in minutes.
BAT	Numerical feature representing the batch of a delivery, i.e., the average number of orders the driver delivered per route.
RAIN	Numerical feature acting as a proxy for the likelihood of rain, derived from information on precipitation intensity and temperature.
available_slots	Numerical feature representing the number of slots made available for booking for that specific shift.
fill_rate	Numerical feature representing the percentage of scheduled slots, from those available.
attendance_rate	Numerical feature representing the percentage of drivers that attended their previously scheduled slot.
permanence_rate	Numerical feature representing the percentage of the shift duration that the attending drivers were online. (for example, a driver that booked a 4-hour shift and stayed online for 3 of those 4 would have a 75% permanence rate).

Table 1 - Feature Set for causality framework. 12 variables were used in this feature set.

Feature selection for the predictability framework (outlined in Table 2) was restricted by the need to keep the feature set similar to the one employed by the benchmark models, for comparison purposes. Despite this, some significant modifications were done to the feature set of the benchmark models. Albeit present on the original benchmark models, features regarding driver's promotion and registration plans were not included, as they are manually generated and ingested by internal teams via custom-made pipelines. Furthermore, they are not very important to the benchmark models, indicating that they could be excluded from the tested models without significant loss of comparability.

A binary feature for holidays was also not included, for the following reasons: 1) there would be very few data points in the dataset with a positive value; 2) the effect of holidays on demand and supply hour varies greatly, depending on the holiday. Due to this, the impact of a holiday on the dependent variable is not capturable by a binary feature.

On the other hand, operational features not originally present in the benchmark models were included, informed by domain knowledge and the preliminary results from the causality framework: ORD, BAT, and EDT. OCC was also included, due to its relationship with SH.

All "lag7" features represent the given information, with a lag of 7 days - i.e., the values for the features of the same shift in the week prior. Due to time constraints, little experimentation was carried out regarding different time lags. We chose to utilize the same time lag as the benchmark models, leveraging the business insight gleaned from that previous research. Recommendations and future steps based on this topic are explored in the limitations section.

Feature	Description
weekday	Categorical feature representing the day of the week. (Monday-Sunday)
logistic_region	Categorical feature representing the logistical region of the observation.
shift	Categorical feature representing the shift of the record (morning, lunch, afternoon, dinner, night)
ORD_lag7	Numerical feature representing demand, i.e., the number of orders.
SH_lag7	Numerical feature representing the Supply Hour.
EDT_lag7	Numerical feature representing the average expected delivery time of an order, in minutes.
BAT_lag7	Numerical feature representing the batch of a delivery, i.e., the average number of orders the driver delivered per route.
RAIN	Numerical feature acting as a proxy for the likelihood of rain, derived from information on precipitation intensity and temperature.
OCC_lag7	Numerical feature representing the average occupation of the shift. Occupation is defined as the ratio of a driver's worked hours / online hours, i.e., the percentage of time a driver was occupied with the delivery of an order.
available_slots	Numerical feature representing the number of slots made available for booking for that specific shift.
fill_rate_lag7	Numerical feature representing the percentage of scheduled slots, from those available.

attendance_rate_lag7	Numerical feature representing the percentage of drivers that attended their previously scheduled slot.
permanence_rate_lag7	Numerical feature representing the percentage of the shift duration that the attending drivers were online. (for example, a driver that booked a 4-hour shift and stayed online for 3 of those 4 would have a 75% permanence rate).

Table 2 - Feature Set for the predictability framework. 13 variables were used in this feature set.

3.5. Modeling

Given the objective of discovery outlined in section 3.1. regarding the project's framework, the modelling choices were driven by the need to be simple in application and increase diversity in scope.

The following models were used: Linear Models, kNN (K-Nearest Neighbors) (Taunk et al., 2019), and Ensemble Decision Trees (Random Forest Regressor (Breiman, 2001), Gradient Boosting Trees (see Natekin & Knoll, 2013). Additionally, a dummy model was applied to get an error baseline, for comparison.

This allowed for testing with a varied range of models.

- Linear Models were used exclusively for the causality framework, due to their usefulness for fitting data to easy-to-interpret parametric models, and suitability for causal inference.
- kNN were used due to their fast training and testing times, as well as good performance in datasets with clusters of similar data points, as this one is for different categories such as logistical region and shift.
- The Decision Trees were chosen for their relative ease of use, and generally good results in regression problems with multi-categorical data such as these. Boosting is also generally appropriate for problems with small datasets, although ideally mitigating those issues should not be done through the choice of a model.

The training data was split using the sklearn package, with a test split of 0.1 of the dataset. Models were trained on operational and booking data, with Supply Hour as the dependent variable.

A 10-fold cross-validation was used in conjunction with GridsearchCV for hyperparameter tuning.

Using out-of-time validation was excluded due to data scarcity and significant gaps in the historical data.

3.6. Evaluation and Comparison

For the error of the benchmark model to be comparable to those trained in the scope of this project, the scale and granularity of the model need to be the same. Due to the architecture of the databases used, the range and duration of a shift differ in both models (e.g., the morning shift in the benchmark models' dataset may range from 9 to 12 PM whereas the booking morning shift ranges from 10 to 1 PM), and so scale and granularity equivalence could not be guaranteed. However, the difference in scale was understood to be minimal enough to still allow for preliminary comparisons.

3.6.1. Metrics

MSE (Mean Squared Error) was used for model training. For evaluation, RMSE (Root Mean Squared Error) and R2 (Coefficient of Determination) were additionally used as evaluation metrics. This choice is supported by the need for highly interpretable statistics, for the presentation of results to the business team[6].

3.6.2. Feature Importance

To quantify the feature importance of all used variables, permutation importance (*4.2. Permutation Feature Importance — Scikit-Learn 1.1.2 Documentation*, n.d.) was used. The importance is represented by the decrease in R2 from shuffling the feature. This metric is model-agnostic and more easily interpretable. Should a future model perform better in this prediction problem, the results can still be compared to the present models to quantify the improvement in what concerns booking variables.

3.6.3. Significance and Coefficient Analysis

The OLS module from the statsmodels packages was used to fit the causality framework dataset to extract several statistics related to causal inference and garner complementary insights.

To test the overall significance of fitted linear models with and without booking variables, an F-test was conducted, by using the available tools from the aforementioned statsmodels package. This tests if the added explanatory power of an unrestricted model (i.e., with the booking variables) is significant enough in contrast to a restricted model.

[6] To this end, RMSE is especially important; it represents the same unit scale as the dependent variable, facilitating its analysis.

4. RESULTS AND DISCUSSION

In this section, we present the results of the tested models. For each framework, we present the model test scores, the corresponding R2 scores, and the feature importances of the best-performing model.

We then discuss some of the most significant results and their implications in the context of the research question.

4.1. Causality Framework Results

Figure 7 and 8 present the model and R2 scores for the causality framework:

Figure 7 - Model Scores (Causality Framework). For each model, both train MSE (mean squared error) and test MSE are shown. The red label shows the specific value of the test MSE. Models are ordered from best to worst (from smallest to larges error). The Ensemble Tree Models (Gradient Boosting Regressor and Random Forest Regressor) achieved the best score, with a significant difference between the two. The K-nearest neighbors model achieved a score very similar to the RF Regressor. The best linear model performed worse than the other three models and better than the Dummy Model, with a baseline error of over 100’000.

Figure 8 - R2 Scores (Causality Framework). All models presented exceptionally high goodness of fit, with R2 above 0.9. The ranking of the models is the same as the one in Figure 7, despite the KNN model tieing with RF regressor for goodness of fit.

Figure 9 presents the feature importances of the best performing model, GB Regressor.

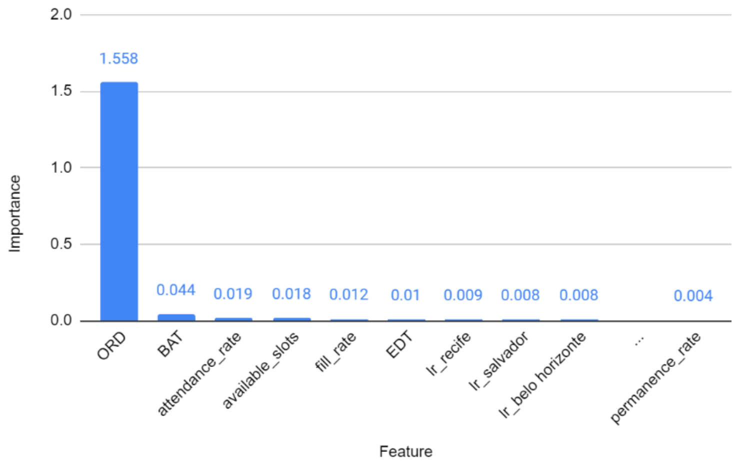

Figure 9 - Feature Importances (Causality Framework). ORD is the most important feature for the model, by a factor of over 35 in relation to the next most important feature, BAT. Attendance rate, available slots, and fill rate are the next most important features, followed by EDT and logistical region categories. The other booking-related feature, permanence rate, has one of the smallest importances.

The feature importances show that the feature regarding the number of orders (ORD) is extremely important. Moreover, the R2 test scores indicate that most models are well suited to explain the observed variability of the dependent variable, with values very close to 1.

Due to their size, the OLS results of the fitted linear models using statsmodels API are presented in Appendix B.

The results show most chosen features (including booking features) have p-values of 0, hence being individually statistically significant.

The computations necessary for the F-test are presented in the following table (the formula for the F-test can be found in appendix C):

Calculation Component	*Value*
SSR - Restricted Model	293519443.89
SSR - Unrestricted Model	270818564.12
N (number of instances)	2973
K (number of explanatory variables)	34
q (degrees of freedom)	4
F-statistic value	61.59
p-value	0

Table 3 - Relevant Calculations for the F-test

As the p-value is zero, we reject the null hypothesis that the restricted model is the correct one; the unrestricted model has a significantly better fit for the data.

4.2. Predictability Framework Results

The model test and R2 scores for the predictability framework are presented in Figures 10 and 11.

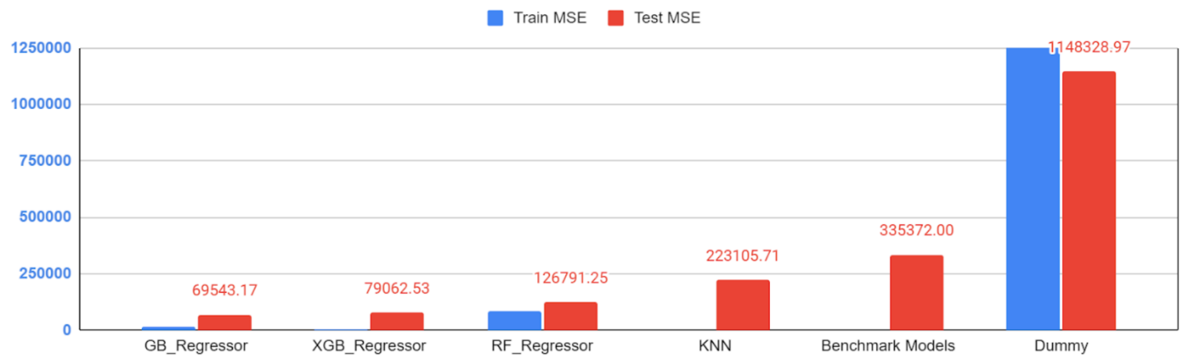

Figure 10 - Model Scores (Predictability Framework). For each model, both train MSE (mean squared error) and test MSE are shown. The red label shows the specific value of the test MSE. Models are ordered from best to worst (from smallest to larges error). The Ensemble Tree Models (Gradient Boosting Regressor, Extreme Gradient Boosting Regressor and Random Forest Regressor) achieved the best scores, albeit GB Regressor model achieved the best score with a significant margin. The K-nearest neighbors model's score is worse by a significant margin. Despite being based on RF regression, the benchmark models have a larger Test MSE than all other models.

Figure 11 - R2 Scores (Predictability Framework). The order of the presented models is the same as Figure 10. All models presented high goodness of fit to their data. The KNN model presented the worst goodness of fit, at 0.8. The benchmark models achieved a R2 score only 0.2 points below the RF regressor model.

Figure 12 depicts the feature importances of the best performing model (GB Regressor).

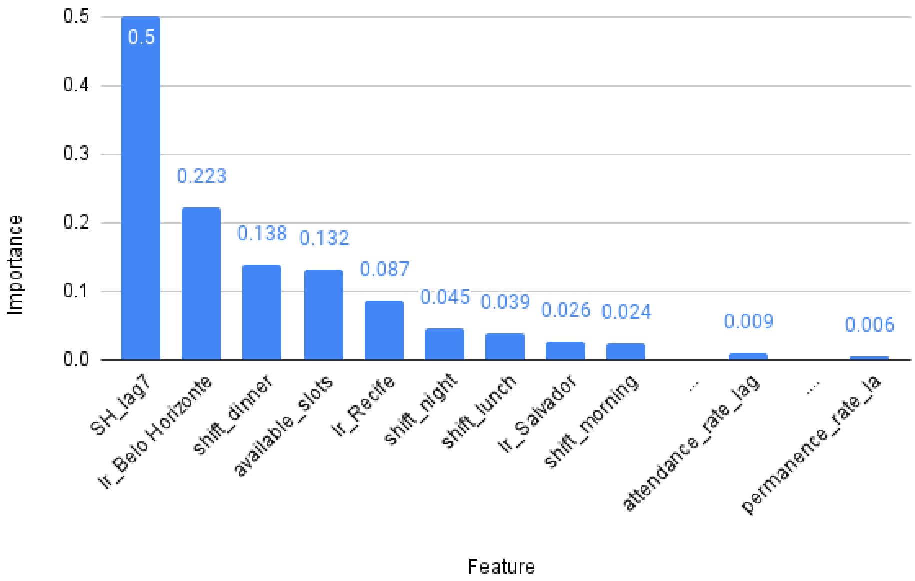

Figure 12 - Feature Importances (Predictability Framework) of the GB regressor model. Supply Hour (lagged) is the most important variable, followed by categories of logistical region and shift. Available slots ranks as the third most important variable, followed by other categories of both logistical region and shift. No other operational or booking-related variable achieves importances above 0.02.

As with the benchmark models, the most important feature is the lagged Supply Hour. Of all Booking features, only the number of available slots appears in the ten most important features.

4.3. Discussion

4.3.1. Operational Control

Domain knowledge indicates there is a causal relationship between orders and Supply Hour: drivers will stay online longer if they keep receiving orders, and there is shared community knowledge about demand peaks. Moreover, additional pay per route is offered to the driver when situations of undersupply occur, further incentivizing supply to rise to meet the demand. The results, in the context of the causality framework, corroborate this knowledge. They also provide a comparable quantification of its impact, specifically concerning booking variables.

We posited that, should the booking system exert any significant control over the operation, a top-performing model would find the respective variables important. In other words, if there would be enough cases where the imposed limitations of slots affected the available supply, as depicted in the literature review, we would not observe feature importances so close to 0. Moreover, in the fitted linear models, the coefficient differences between ORD and available slots indicate that, despite being statistically significant, the extent of that control is very limited.

The observed results are indicative that while some form of causal relationship exists between the defined number of slots and supply hour, the relationship is still inelastic.

Since this text was written, the business team has conducted a few narrow-scoped quasi-experiments (such as removing the limit on slots for specific shifts), with no significant observed effects on the amount of Supply Hour available, corroborating the findings of this research.

4.3.2. Predictability Framework

Regarding predictability, we found that out of all booking-related features, the number of available slots is the most important - and arguably, the only one. Importance-wise, it outperforms the additional operational features (ORD, EDT, BAT, OCC). Additionally, preliminary testing involving a pruned feature set (removing all features except categorical, lagged SH, available slots, and lagged ORD) revealed slightly better scores for all tested models (see appendix D).

In this sense, while technically it could be said that booking-related variables are relevant, and therefore the booking system increases the performance of predictive models, the effect appears marginal. Moreover, we should consider the underlying process that generates the value, and whether it can be attributed to, or dissociated from, the booking system. In practice, the number of available slots is calculated by taking the predicted number of orders (demand) and then adjusting that number for a predefined level of occupation and UTR (based on predicted values of attendance, permanence, and fill rate[7]). Additionally, we should take the insights gleaned from feature importances of the feature set used for the causality framework into account, where same-day orders had extremely high importance. When the orders feature is excluded from the feature set, the available_slots feature, which includes information on same-day orders (even if just a prediction) gains importance. Due to their correlations, the hypothesis is that available_slots contains information relating to same-day orders. In that sense, we posit that integrating a feature related to predicted orders may benefit the benchmark operational models (regardless of the existence of the booking system). The extent of this benefit, however, is unclear.

Regarding the other variables (lagged permanence, attendance, and fill rate), we can once again look at the causality framework results to extract plausible hypotheses. One hypothesis is that these features could act as proxies for the non-lagged features. Whereas significant in the causality framework, the importance for these lagged features are some of the lowest ranked for all tested models, implying that, for the chosen lag, they are not good predictors of the actual value.

[7] Despite possible issues with multicolinearity, the models nonetheless performed better by including the mentioned features in addition to available slots.

5. CONCLUSION

To investigate the baseline assumptions for the booking system, and to answer the research question, ‘Does the Booking system increase the predictability of Supply at iFood?’, we employed data-centered and theory-driven approaches to its analysis.

It is understood the results are not conclusive, but merely indicative. They nonetheless represent strong foundational research on the subject, from which several future steps can be proposed (see Section 6).

We constructed two frameworks of analysis: one to examine the causal relationships between booking variables and supply (measured in supply hours), and one to compare the performance of predictive models equipped with booking variables against the performance of already existing operational models.

With the modelling and testing performed in the first framework, we extracted relevant insights regarding the influence of Booking in supply hours. Given the findings outlined in the literature review (1) that drivers value autonomy of labor hours over possible gains, and 2) they are resistant to systems in which they feel forced to work at certain times), the results of the applied linear regressions, F-test, and feature importances, we inferred that while the booking system is capable of some influence of driver’s behavior via definition of number of available slots, this effect is overwhelmingly small in comparison to the ORD feature (demand). In other words, we posit that supply elasticity is low with regard to gains, and that the booking system will be incapable of influencing driver’s choices of slots without severely limiting their autonomy.

On the topic of predictability (the second framework), there is evidence that suggests the Booking system increases the visibility of drivers’ intentions regarding their activity, which can be leveraged to increase the predictive power of the company’s forecasting models. This was indicated by the performance of the applied models in relation to the benchmark. The feature importances also indicate the value of available_slots, which ranked higher than all other operational features (except the lagged variable for supply hour).

We also discussed that it is uncertain if these intentions are inherent to the driver’s underlying habits, or if they are induced by the Booking system (therefore, artificially created). Given the points discussed in the previous section, we posit they are inherent and not induced.

To summarize, we posited that while the answer to the research question is “yes”, there is indication that the effect is very small and further research, testing, and iteration, is advised. This work has contributed to opening internal avenues through which that can be achieved, which are more thoroughly explored in Section 6.

To conclude, a relevant point should be made regarding driver experience; despite it being out of the scope of the research question, it is relevant to the subject matter of the applied research in the context of a professional internship. The discussion of the pros and cons of the Booking system revolves around the dichotomy of Control vs Autonomy. While the system theoretically provides companies levers through which to exert control over the working hours of drivers, the subsequent degradation of drivers’ autonomy can lead to driver dissatisfaction and churn. Ideally, there would be a balance between these two elements that allows the company to ensure quality service on the most critical shifts while retaining a positive driver experience. To this end, this research contributes with powerful insights that uncover paths to this balance and avoid the very real pitfalls outlined in the literature review.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

In this section, we discuss the constraints and obstacles that limited the scope of the research, its impact on the current research, and propose solutions to mitigate and address them in future ventures on the topic.

6.1. DATA SCARCITY

The most relevant limitation of this research is the scarcity of data. Any model trained and tested in such a restricted scope (both temporally and spatially) is expected to not be generalizable. This limitation is compounded by the imbalance of logistical regions in the dataset. With fourteen different logistical regions, their representation ranges from just 3% of the dataset to over 10%. The effect is visible in the results of the feature importances in the predictability framework, where the features corresponding to logistical regions, or shifts, are highly ranked. The scarcity of data will in some cases confine the range of the target variable in less represented data points, leading to models overfitting. This issue is mitigated by using boosting algorithms, which cope well with small or imbalanced datasets. Another possibility is to merge less-represented categories (doing so for the current project was deemed too time-consuming, and the logistical regions marked as most important are those who would not get merged due to their size). Considering the vast number of logistical regions, this will certainly be necessary to achieve high-performing models. A suggested approach is to merge less frequent categories with unsupervised learning algorithms (to group categories containing similar information and relations).

6.2. CAUSAL INFERENCE DISCOVERY

Including issues of data scarcity, other obstacles exist to causal inference in this project. For example, given that the sampled data are historical, it is likely that instances are serially correlated, and not truly independent. This affects the reliability of the conducted F-test, as well as the estimation of the parameters in the linear models.

Additionally, the obtained results are not reliable in quantifying the impact of the introduction of the booking system on operational control. For example, the naive approach would be to take the estimated parameter of available slots and interpret it as a measure of gained control. But as discussed, the variable is not truly independent of external and uncontrolled factors (i.e., orders), which means the level of exerted influence is unclear and very likely to remain so.

One approach we considered was that, if the booking variables had a positive impact on predictability, that would be reflected on the scores of models trained with data from before and after introducing Booking. This approach is non-robust because even when considering the same logistical region before and after the introduction of booking, we cannot guarantee that other factors are held constant (e.g., seasonality), which makes attributions of cause to be less reliable.

Instead, we suggest an experimental approach, via the identification and comparison of models to control groups; in this case, compare similar logistical regions where the only difference is the existence of the booking system. A viable alternative to this identification is to apply synthetic control techniques (Abadie, 2021). This method permits the construction of a weighted combination of control groups to which a treatment effect can be compared to. This, however, would require significant rework of model specifications, as the impact we want to quantify is not on the dependent variable (supply hour), but rather on the parameter of an independent variable.

6.3. Time Series Forecasting

In general, this type of predictive problem fits with the use cases for time series forecasting. However, for the purpose of this research, we excluded the usage of time series modeling early on, for the following reasons:

- Data Scarcity: data was available only for a short period of time, deemed not enough for long-term patterns to stabilize or be visible. Moreover, data discrepancies generated periodical gaps in historical data.
- Exogenous features: the dataset contains many exogenous features, which models such as Gradient Boosting Trees can handle better than traditional Time Series models.
- Seasonality: Optimizing a time series model requires extensive experimentation to find an appropriate time lag. In this project's case, such a time lag had been previously determined; moreover, applying a different time lag would make comparison with the benchmark models more difficult.

Considering these points, the application of time series modeling to a forecasting problem with booking features is indicated for future research, with emphasis on developing robust pipelines to guarantee the generation of reliable and continuous historical data.

6.4. Refinement of Operational Benchmark Models

As discussed in a previous section, the findings of this research are obscured by what can be attributed to the booking system and what can be attributed to the suboptimization of the benchmark models. Regardless of the underlying reason, the feature available_slots increases the predictive power of these models, and so its integration is suggested. In regions that do not yet have Booking, integrating a feature of same-day predicted demand is suggested. This will allow for further investigation on effect attribution, by comparing the performance and importance of the respective features.

Additionally, we discovered some preliminary significance of fill rate, attendance rate, and permanence rate. In a previous section, we discussed the importance of available_slots in the context of the observed results between the two frameworks. Applying the same process to the other booking variables, we note that while significant for the fitted linear models under the causality framework, the lagged variables are decidedly less relevant when applied for prediction problems. However, the used time lag was only a placeholder and unlikely to be predictive. In the future, different time lags should be tested for these features; as well as differently engineered features based on that information.

On the practical side, a concurrent suggestion was made to the business team: incentivize constant levels of attendance, permanence, and fill rates. For context, recall that the underlying formula relating these features (*Booking SH = Available Slots * Fill Rate * Attendance Rate * Permanence Rate * SH per slot)* simply represents their product. When constant, the features could be removed from the dataset. In other words, by incentivizing habitual, assertive behaviors in drivers, previously dispersed information is concentrated on one non-lagged feature (available_slots).

7. REFERENCES

4.2. Permutation feature importance — scikit-learn 1.1.2 documentation. (n.d.). Retrieved August 29, 2022, from https://scikit-learn.org/stable/modules/permutation_importance.html

Abadie, A. (n.d.). *Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects*. https://doi.org/10.1257/jel.20191450

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Drahokoupil, J., & Piasna, A. (2019). Work in the Platform Economy: Deliveroo Riders in Belgium and the SMart Arrangement. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3316133

Gandini, A. (2018). Labour process theory and the gig economy: *Https://Doi.Org/10.1177/0018726718790002*, *72*(6), 1039–1056. https://doi.org/10.1177/0018726718790002

Griesbach, K., Reich, A., Elliott-Negri, L., & Milkman, R. (2019). Algorithmic Control in Platform Food Delivery Work: *Https://Doi.Org/10.1177/2378023119870041*, *5*, 237802311987004. https://doi.org/10.1177/2378023119870041

Heiland, H. (n.d.). *Neither timeless, nor placeless: Control of food delivery gig work via place-based working time regimes*. https://doi.org/10.1177/00187267211025283

Ivanova, M., Bronowicka, J., Kocher, E., & Degner, A. (2018). *Foodora and Deliveroo: The App as a Boss? Control and Autonomy in App-Based Management – The Case of Food Delivery Riders*. https://creativecommons.org/licenses/by/4.0/de/legalcodewww.econstor.eu

Malik, R., Visvizi, A., & Skrzek-lubasińska, M. (2021). The Gig Economy: Current Issues, the Debate, and the New Avenues of Research. *Sustainability 2021, Vol. 13, Page 5023*, *13*(9), 5023. https://doi.org/10.3390/SU13095023

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(DEC). https://doi.org/10.3389/FNBOT.2013.00021/ABSTRACT

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/J.PROCS.2021.01.199

Stewart, P., Shanahan, G., & Smith, M. (2020). Individualism and Collectivism at Work in an Era of Deindustrialization: Work Narratives of Food Delivery Couriers in the Platform Economy. *Frontiers in Sociology*, *5*, 49. https://doi.org/10.3389/FSOC.2020.00049/BIBTEX

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747

Tony Liu, F., Ming Ting, K., & Zhou, Z.-H. (2008). *Isolation Forest*.

Veen, A., Barratt, T., & Goods, C. (2019). Platform-Capital's 'App-etite' for Control: A Labour Process Analysis of Food-Delivery Work in Australia: *Https://Doi.Org/10.1177/0950017019836911*, *34*(3), 388–406. https://doi.org/10.1177/0950017019836911

Zhong, Y., Lin, Z., Zhou, Y. W., Cheng, T. C. E., & Lin, X. (2019). Matching supply and demand on ride-sharing platforms with permanent agents and competition. *International Journal of Production Economics*, *218*, 363–374. https://doi.org/10.1016/J.IJPE.2019.07.009

8. APPENDIX

8.1. APPENDIX A

Bar plot of available records per logistical region

Figure A. 1 - Number of availble records in the dataset, by logistical region. Logistical regions are ordered from most populatred to less populated. Labels for each bar represent the absolute number of records and their proportion in the dataset. The top 5 regions represent about 50% of all records of the dataset.

8.2. Appendix B

Figure B.1. presents the

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.942
Model:                            OLS   Adj. R-squared:                  0.941
Method:                 Least Squares   F-statistic:                     1441.
Date:                Sat, 20 Aug 2022   Prob (F-statistic):               0.00
Time:                        16:24:46   Log-Likelihood:                -21194.
No. Observations:                2973   AIC:                         4.246e+04
Df Residuals:                    2939   BIC:                         4.266e+04
Df Model:                          33
Covariance Type:            nonrobust
=================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
weekday_1                        132.5487     20.664      6.415      0.000      92.032     173.066
weekday_2                        136.8083     20.718      6.603      0.000      96.185     177.432
weekday_3                        130.3988     20.773      6.277      0.000      89.668     171.129
weekday_4                        190.6892     21.883      8.714      0.000     147.782     233.596
weekday_5                         53.3031     21.518      2.477      0.013      11.110      95.496
weekday_6                        -51.3931     22.508     -2.283      0.022     -95.526      -7.260
logistic_region_Belo Horizonte   755.3595     48.318     15.633      0.000     660.619     850.100
logistic_region_Campinas          92.6033     40.578      2.282      0.023      13.039     172.168
logistic_region_Florianopolis   -249.5678     40.208     -6.207      0.000    -328.407    -170.728
logistic_region_Fortaleza        172.9439     34.870      4.960      0.000     104.571     241.317
logistic_region_Goiania          214.3104     29.513      7.262      0.000     156.443     272.178
logistic_region_Joao Pessoa     -384.4552     39.903     -9.635      0.000    -462.696    -306.215
logistic_region_Jundiai         -379.5906     44.887     -8.456      0.000    -467.605    -291.577
logistic_region_Maceio          -138.6430     30.562     -4.536      0.000    -198.569     -78.717
logistic_region_Porto Alegre      40.9994     28.147      1.457      0.145     -14.191      96.189
logistic_region_Recife           849.4062     40.966     20.735      0.000     769.082     929.731
logistic_region_Ribeirao Preto  -199.0065     25.149     -7.913      0.000    -248.319    -149.694
logistic_region_Salvador         366.7380     33.608     10.912      0.000     300.840     432.636
logistic_region_Sorocaba        -231.8775     28.349     -8.179      0.000    -287.464    -176.291
logistic_region_Uberlandia      -166.6117     38.948     -4.278      0.000    -242.980     -90.243
shift_dinner                     233.6800     21.710     10.764      0.000     191.111     276.249
shift_lunch                       47.2199     20.968      2.252      0.024       6.106      88.334
shift_morning                   -216.4437     21.478    -10.078      0.000    -258.557    -174.331
shift_night                     -256.8796     24.205    -10.613      0.000    -304.340    -209.419
ORD                              905.8766     13.679     66.225      0.000     879.056     932.698
EDT                               74.1857     13.360      5.553      0.000      47.989     100.382
RDT                              -56.1164     12.759     -4.398      0.000     -81.133     -31.099
BAT                             -147.4722     10.685    -13.802      0.000    -168.423    -126.522
available_slots                   58.1864      9.044      6.434      0.000      40.453      75.920
fill_rate                         55.9944      9.706      5.769      0.000      36.963      75.026
attendance_rate                   96.0454      8.806     10.907      0.000      78.780     113.311
permanence_rate                   41.2163      8.901      4.630      0.000      23.763      58.670
RAIN                             -42.4561      6.310     -6.729      0.000     -54.828     -30.084
const                           1457.8531     27.006     53.982      0.000    1404.900    1510.806
==============================================================================
Omnibus:                      777.612   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9908.728
Skew:                           0.875   Prob(JB):                         0.00
Kurtosis:                      11.771   Cond. No.                         26.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure C. 1 - OLS Regression results, as outputted by the statsmodels package, of fitted linear models (causality framework)

8.3. Appendix C

Formula for the f-test of overall significance

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n-k-1)} \sim F_{q,n-k-1}$$

where q is equal to the number of restrictions, SSR is the residual sum of squares from the restricted and unrestricted models, respectively, n is the sample size, and k is the number of explanatory variables in the unrestricted model.

8.4. Appendix D

Comparison of model scores for pruned feature set (predictability)

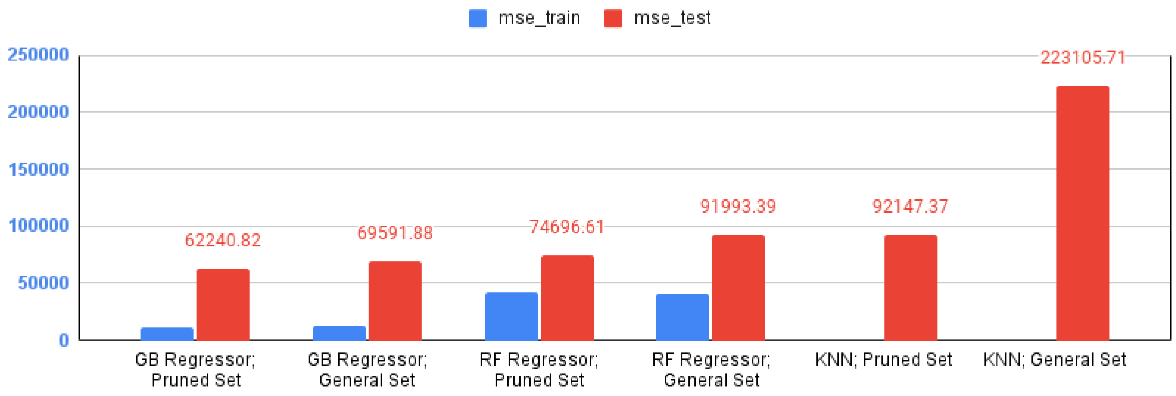

Figure D. 1 – Comparison between the general set used and a pruned feature set in the applied models of the predictability framework. Utilizing a pruned feature set shows significant improvement in model errors, with the K-nearest neighbors model benefitting the most from the exclusion of unimportant variables.

Table D.1 describes the two feature sets.

Name	Features	Number of Features
General	Weekday, logistic_region, shift ORD_lag7, SH_lag7, attendance_rate_lag7, permanence_rate_lag7, available_slots, fill_rate_lag7, BAT_lag7, EDT_lag7, RAIN, OCC_lag7	13
Pruned	Weekday, logistic_region, shift ORD_lag7, SH_lag7, available_slots	6

Table D. 1 – Feature comparison between the two sets. The "general" name is descriptive of the feature set being the same as the one used in the main body of work for the predictability framework. The pruned feature set has 7 features less than the general one.

All models underwent the same pre-processing and training pipeline, with corresponding gridsearch-based parameter tuning (same parameter spaces for both sets). Due to this, error scores for some models differ from the ones presented in Section 4, none of which impact the obtained results and subsequent discussion.

Overall, the results highlight the impact of a well-applied feature selection stage in predictive modelling tasks.