Decoding Carrier Preferences in Digital Freight: A Predictive Machine Learning Analysis

by

Perry Falk

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE OR MASTER OF ENGINEERING IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Signature of Author: ______________________________
Department of Supply Chain Management
May 10, 2024

Certified by: ______________________________
David HC Correll, PhD
Co-Director, MIT FreightLab
Capstone Advisor

Accepted by: ______________________________
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Decoding Carrier Preferences in Digital Freight: A Predictive Machine Learning Analysis

by

Perry Falk

Submitted to the Program in Supply Chain Management
on May 12, 2024 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science or Master of Engineering in Supply Chain Management

ABSTRACT

The freight brokerage industry is at a pivotal juncture, with digital platforms reshaping market dynamics and carrier preferences. This capstone project, undertaken in partnership with Nolan Transportation Group (NTG), employs a predictive machine learning model to decode and understand these evolving preferences. The study leverages a dataset comprising nearly 2 million brokerage transactions, enriched with comprehensive feature engineering, to model the likelihood of digital vs. traditional booking methods. The research uses advanced machine learning algorithms, especially Gradient Boosting with XGBoost, to identify key shipment characteristics that influence carriers' digital booking decision untangling the complex interplay of shipment characteristics that influence digital booking decisions. Central to these findings is the pivotal role of the time a load remains available on digital platforms in determining its likelihood of being digitally booked. The analysis underscores a critical insight: the probability of a load being booked digitally diminishes significantly with time, highlighting a narrow window for digital engagement. This discovery has valuable operational implications, suggesting a strategic shift towards minimizing internal competition for loads in the period of initial listing, thereby enhancing the effectiveness of digital channels. By offering a nuanced understanding of the temporal dynamics at play in digital freight booking, this research provides actionable strategies for fostering digital adoption and optimizing brokerage operations in the digital age. Through this lens, the study not only contributes to academic discourse but also equips industry practitioners with the insights needed to navigate the evolving landscape of freight brokerage.

Capstone Advisor: David HC Correll, PhD
Title: Co-Director, MIT FreightLab

ACKNOWLEDGMENTS

I want to first thank my wonderful wife, Brooke, and my daughter, Porter, for their support during this incredible journey. I would not be here without them, and I certainly would not have completed this task without their help. And of course, thanks to Bear and Weber.

Next, I want to express my appreciation to Dave Correll, who was a fantastic advisor for this work. After nearly a year of calls, meetings, advice, and revisions, I am so proud of the work we did together. The entire team of professors, administrators, and support within MIT CTL have been so supportive during this entire process.

Finally, I want to give a special thanks to my sponsor company (and employer), NTG. The company's support in this research was invaluable. But more valuable than that was NTG's support and confidence in me to undertake this program.

TABLE OF CONTENTS

1. INTRODUCTION

The freight brokerage industry has been disrupted by digital platforms, presenting opportunities and challenges. To adapt to the changing landscape, it's crucial to understand carrier preferences in both digital and traditional contexts. In partnership with Nolan Transportation Group (NTG), this research aims to uncover these preferences and refine technology to lead the digital brokerage revolution.

Freight brokerage in North America is a highly competitive and fragmented marketplace, with hundreds of brokerage firms competing for business from thousands of shippers and for partnerships with thousands more carriers. Historically, success in this domain has been achieved using legacy tools – load boards, phones, email, fax machines – combined with human capital, experience, and solid partnerships. These tools and techniques served as the bedrock of the industry for years.

The status quo started to change between 2015 and 2017 when the landscape was disrupted by the emergence of “digital freight brokers” like Convoy and Uber Freight (Balakrishnan, 2017; Holland, 2022). Their entry heralded a shift: Once a tool, technology became a competitive differentiator. These firms promised to use new technology to eliminate the need for traditional brokerages to be the “middleman” between shippers and carriers. As the allure of these digital platforms grew and the threat of this disintermediation loomed, the industry saw the need to redefine its operational strategies. Technological advancements' swift pace promised efficiency but rendered many established practices obsolete. These traditional best practices and 'rules of thumb' now faced a challenge: adapting to the rapidly changing digital landscape.

NTG has embraced this digital shift, recognizing this evolution and its profound implications. They began their digital transformation in 2020, establishing their place as a hybrid digital brokerage (BOSS Editorial, n.d.), combining traditional brokerage with modern technology. Their digital booking platform, Beon Carrier, is a testament to the changing tides, witnessing exponential growth and signaling a clear shift in carrier preferences. But with this growth comes a challenge and an opportunity: to decipher what attracts carriers to these digital platforms over traditional avenues. NTG seeks to uncover these underlying preferences to refine its technology suite and lead the digital brokerage revolution.

As legacy freight brokers began to digitize and adapt, a significant knowledge gap emerged: understanding the nuanced drivers behind carrier decisions in a digital versus traditional context. In partnership with NTG, this research fills this gap, providing novel insight into how booking times relate to digital success. My findings show that digital shipments tend to be booked more quickly than shipments booked through traditional channels. Furthermore, I propose strategies for how brokerages like NTG can better grow their digital booking platforms by protecting the initial booking period and reducing the internal cannibalization of high-potential digital shipments.

The urgency of this endeavor is fueled by the industry's "uberization," a trend that accentuates the need to optimize digital platforms for load allocation. In response to this need, this capstone unravels the complexities of carrier preferences amidst the digital revolution to equip the industry with insights that can redefine the future of freight brokerage.

1.1 Digital Booking vs Traditional Booking

Matching shipments with carriers has long been the cornerstone of freight brokerage operations. Non-asset intermediaries like NTG rely on these connects as their entire value proposition. When a shipment is tendered to a broker, they must match this “uncovered” load with a carrier. Traditionally, this matchmaking was facilitated through interpersonal communication and analog tools. Brokers relied heavily on networks built through phone calls, emails, and even fax machines to negotiate deals between shippers and carriers. Load boards are often the primary tool to build these networks and make these connections. DAT Freight & Analytics, commonly known as “DAT,” is the market-leading load board (Granato, 2019).

This manual approach required significant human intervention, with brokers acting as intermediaries, leveraging load boards, their expertise, and relationships to find the best match for each shipment. The efficacy of this method was largely dependent on the broker's ability to leverage these tools, personal connections, and knowledge of the market, often leading to a time-consuming and costly process that requires extensive personnel and operational precision to execute effectively.

Digital booking platforms have revolutionized this traditional approach by leveraging technology to streamline matchmaking. Digital booking is a self-service process in which carriers select their desired shipment without human intervention. This simple but powerful technology reduces manual labor and facilitates faster and more effective matching.

NTG’s Beon Carrier platform is supported by sophisticated algorithms that analyze vast amounts of data on carrier preferences, shipment details, and historical performance to quickly identify the most suitable matches. These matches are then automatically sent to prospective carriers, speeding up the process, increasing accuracy, and reducing operational costs. Generally, digital booking platforms offer dynamic pricing and enhanced transparency, benefits that traditional methods struggle to match.

This shift towards digital booking reflects broader logistics and supply chain management trends towards automation and data-driven decision-making. By reducing reliance on human brokers and analog tools, digital platforms are changing how shipments are matched with carriers and redefining the competitive landscape of the freight brokerage industry. This digital transformation presents opportunities and challenges, as companies must navigate integrating new technologies with traditional operations to stay relevant in a rapidly evolving market. Understanding the drivers behind carriers' preferences for digital over traditional booking methods is crucial for any brokerage aiming to succeed in this new digital era.

1.2 Problem Statement and Research Questions

NTG’s objective is to understand which shipment characteristics offered on their digital carrier platform have the most significant impact on a carrier's preference. They believe that by understanding what carriers are looking for, they can better tailor their freight offers to increase digital bookings. Increased digital bookings will lead to lower operating costs and a competitive advantage. Understanding these preferences is crucial to long-term business success.

The key questions to answer:

1. What shipment characteristics have the most significant impact on digital booking success?
2. Is a shipment more or less likely to be booked digitally?
3. What actions can brokers take to improve digital booking success?

Understanding why the new booking platform is succeeding compared to traditional booking methods will help inform the strategy for continued growth. With this understanding, brokerages can predict which shipments are more likely to be digitally booked.

Furthermore, NTG wants a proof-of-concept model that can accurately predict whether a new shipment will be booked digitally.

1.3 Hypothesis

I posit that a machine learning-centric approach can determine the intrinsic factors governing carrier preferences in a digitalized environment. By understanding what drives these preferences, it will be possible to predict behavior and then enhance offerings and allocation processes, reducing inefficiencies. Thus, a proof-of-concept model to operationalize these insights will be delivered.

1.4 Scope: Project Goals and Outcomes

The essence of this research project lies in its dedication to bridging a critical knowledge gap within the freight brokerage industry. The project, through a meticulously structured approach utilizing machine learning and data analytics, helps to decode carriers' nuanced preferences when choosing between digital and traditional freight brokerage services, leveraging new insights for practical application within brokerage businesses.

Simply put, the core goal was to empirically understand carrier preferences when faced with digital load offers. The outcomes included:

- **Improved and Novel Insights of Carrier Preferences:** The heart of this research provides a clearer understanding which factors most influence carrier decisions between digital vs traditional load offerings while offering novel insight into these behaviors.
- **Predictive Model for Carrier Behavior:** Leveraging machine learning, a proof-of-concept model was created to predict whether a new shipment will likely be booked digitally or through traditional analog channels.
- **Actionable Insights for Brokers:** By unraveling complexity and decoding the new "rules of thumb," I deliver actionable insights into operational changes that will result in better digital booking outcomes.

These project outcomes deliver valuable insight to both the sponsoring company, and other brokerages who are seeking similar improvements to digital engagement.

1.5 Tangible Benefits for Freight Brokerages

Implementing the recommendations promises a wide range of tangible benefits to both NTG and the broader brokerage community. For NTG, these benefits will fortify its leading position in the dynamic realm of freight brokerage.

- **Enhanced Decision-Making:** The research identifies the nuanced preferences of carriers in choosing digital platforms over traditional methods. A deeper understanding of these predilections will enable brokerages to develop more informed and strategic decision-making processes that are attuned to carrier needs and industry trends.
- **Optimized Technological Investments:** By determining what specifically attracts carriers to digital booking platforms, brokers can prudently channel investments into technological advances that further this attraction. They can enhance the features that carriers most prefer, thereby ensuring the technology suite is not only robust but also precisely aligned with user expectations and requirements.

- **Increased Platform Adoption:** Tailoring the digital platforms, like NTG's Beon Carrier, based on an informed understanding of carrier preferences, will likely fuel increased adoption rates. Enhanced user satisfaction and utility will make the platform more appealing, encouraging a broader spectrum of carriers to migrate towards digital bookings.
- **Strategic Competitive Positioning:** Armed with a predictive understanding of carrier behaviors and preferences in the evolving digital landscape, brokers who incorporate these insights will be better positioned to cultivate a strategic edge, ensuring adaptability and resilience amidst industry shifts and competitive pressures.
- **Boosted Operational Efficiency:** The recommendations will facilitate the refinement of a brokerage's operational approach, optimizing workflows and processes to echo the evolving digital trends. Such enhancements promise a boost in operational efficiency, aligning resources and efforts more congruently with industry advancements.
- **Direct Cost Savings:** Improved platform engagement will result in lower operating costs. A primary KPI used to measure transaction costs and, therefore, the success of the business and its digital products is Cost Per Load. These findings can have a direct impact on improving this fundamental KPI.
- **Augmented Customer Relationships:** The insights garnered will allow for both improved and aligned relationships with carriers, understanding, and catering to their evolving needs and preferences with greater precision and relevance.
- **Future-Ready Approach:** Implementing the recommendations will facilitate a future-ready approach, enabling firms to proactively navigate the foreseen and unforeseen shifts in the freight brokerage landscape, underpinning sustainable growth and innovation trajectories.

2. STATE OF THE PRACTICE

Freight brokerage digitization has been a topic of recent research. Understanding the industry's current state is an important context for understanding the research and results. Just as important is understanding the context within NTG. While some insights and results have broad applications, because this topic is hyper-focused on the company’s specific platform and outcomes, many results are highly specific to NTG.

2.1 Industry

The freight brokerage industry is evolving rapidly. There are new technologies and large firms leading the way.

2.1.1 Digital Freight Brokers

Spurred by competition from the new Digital Freight Brokers (DFB) like Uber Freight and Convoy, traditional brokers are working to keep pace with new tools and processes. The introduction of digital tools has evolved the way these brokers operate. Digital platforms leverage technology to optimize the freight matching process, providing efficiency and cost-effectiveness.

The authors of *Does the Sharing Economy Technology Disrupt Incumbents? Exploring the Influences of Mobile Digital Freight Matching Platforms on Road Freight Logistics Firms* (Zhou & Wan, 2021) illustrates the transformative impact of such platforms on the industry, disrupting traditional brokerage methods. Similarly, Helguera Sánchez and Hendra Mukti, in their capstone project, highlight the operational efficiencies achieved by these digital platforms (Helguera Sánchez & Hendra Mukti, 2018).

These DFB and Digital Freight Matching (DFM) platforms have reshaped the competitive landscape. Machine learning (ML) models are used to evaluate and understand operational processes in both digital and traditional (“non-digital”) freight brokerages. These models have provided new and deep insights into a new set of tools.

After reviewing dozens of articles and papers, I will reference a few as I dive deeper into the tools being used to understand the current landscape and its challenges.

2.1.2 Machine Learning in Freight Brokerage

Machine learning (ML) has become increasingly crucial in freight brokerage, with applications ranging from freight rate prediction to route optimization. Acocella and Caplice in *Research on Truckload*

Transportation Procurement: A Review, Framework, and Future Research Agenda emphasize the potential of ML in enhancing decision-making and efficiency in freight procurement (Acocella & Caplice, 2023).

Illustrating this importance, Nolan Transportation Group (NTG) currently uses several ML-based models to inform decision-making in their day-to-day operations. For example, they have developed a Dynamic Pricing model to predict spot market trucking prices in real-time. Having accurate spot rates available in an instant provides a competitive advantage to NTG in winning spot market business because of the best price and speed to quote, two of the most critical components.

2.1.3 Digital Freight Matching Platforms

The introduction of digital freight matching platforms has transformed the logistics sector. These platforms have not only streamlined the process of matching freight with carriers but also altered the traditional dynamics of pricing and carrier-shipper interactions. This digital shift has brought about a greater need for real-time data processing and analysis, underlining the importance of ML in managing these new dynamics efficiently (Helguera Sánchez & Hendra Mukti, 2018).

2.1.4 Carrier Preferences and Decision Factors

In the digital freight brokerage landscape, understanding the multifaceted preferences of carriers is critical. These preferences can range from simple factors like preferred routes and rates to more complex considerations such as load types, delivery windows, and shippers' historical performance metrics. Analyzing these factors through ML helps in tailoring digital offerings to meet carrier expectations better, thus enhancing platform adoption and usage (Acocella & Caplice, 2023).

2.1.5 Operational Efficiency and Cost Implications

The digitization of freight brokerage has substantial implications for operational efficiency and cost management. Digital platforms facilitate better asset utilization, streamline administrative tasks, and reduce overhead costs. Furthermore, they enable more strategic allocation of resources, allowing companies to focus on core competencies and value-added services (Zhou & Wan, 2021).

2.1.6 Future Directions and Challenges

As the industry continues to evolve, it faces challenges like integrating advanced digital systems into existing operations and navigating the complexities of market volatility. The balance between digital and traditional brokerage methods remains a key area of exploration. Without a deeper understanding of these systems, companies like NTG are challenged with the balance of risk between investing in these technologies without substantial gain and the risk of missing out on a digital revolution. Gaining deeper insight and understanding is imperative.

2.2 Nolan Transportation Group (NTG)

Nolan Transportation Group, founded by Kevin Nolan in 2005, is currently one of the largest logistics companies in the United States, with over $15 billion of freight under management. NTG's brokerage ranks in the top 10 largest brokerages, with competitors such as C. H. Robinson, TQL, Coyote Logistics and Uber Freight. Beon Carrier is NTG's proprietary digital booking platform. This platform was developed in-house and serves as the marketplace where trucking carriers can bid on and book NTG's available freight. The platform exclusively offers NTG's shipments to the company's network of nearly 100,000 carrier partners. Beon Carrier facilities hundreds of digital bookings each day and has been a growing competitive advantage for NTG.

3. METHODOLOGY

In this section, I will detail the data collection and preparation process and the methodology used in model creation and formulation.

3.1 Data Collection

NTG provided a massive amount of operational data to facilitate this research. The data set included nearly 2,000,000 brokerage transactions over three years. Each row represented a single freight transaction and included up to 40 unique data points detailed in Table 1.

Through personal insight working at NTG and discussions with the company, it became clear that narrowing the vast data history to a more recent subset was required. The Beon Carrier platform is relatively new. It was revamped, rebranded, and relaunched in 2021. During the relaunch phase, the brokers and carrier sales representatives were incentivized to encourage digital bookings on the platform. These incentives helped encourage early adoption and helped the NTG deliver its new tool.

However, this ramp-up period and incentive offers can potentially mislead the model. With that in mind, I chose to narrow it down to a more recent section of data to test the model on. With the promotional period ending, I focused only on the most recent six-month period. This allowed for a large enough sample size to train the model and uncover the insights NTG hoped to find.

Table 1

Transaction data features and descriptions

Feature	Description
LOADID	Unique internal identification number for each shipment
PICKUPDATE	Date of shipment pick up.
BUILDDATE	The date shipment was created in the system.
FIRSTACTIVETIME	Date time when shipment was first available for booking
LASTBOOKEDTIME	Date time when the shipment was booked
PICKUPADDRESS	Street address of shipment pickup location
PICKUPCITY	City of shipment pickup
PICKUPSTATE	State of shipment pickup
PICKUPZIP	Zip code of shipment pickup
PICKUPCLUSTER	Internal cluster code of pickup
PICKUPREGION	Internal region code of pickup

PICKUPINSTRUCTIONS	String of text input by broker sharing important notes and information about the shipment
DELIVERYADDRESS	Street address of shipment delivery location
DELIVERYCITY	City of shipment delivery
DELIVERYSTATE	State of shipment delivery
DELIVERYZIP	Zip code of shipment delivery
DELIVERYCLUSTER	Internal cluster code of delivery
DELIVERYREGION	Internal region code of delivery
DELIVERYINSTRUCTIONS	String of text input by broker sharing important notes and information about the shipment
REQUIREMENTS	Mandatory requirements from a pre-defined list (i.e., Tarps, Pallet Jack, TWIC)
EQUIPMENT	The exact type of trailer needed (ie, 53; Dry Van, 48' Dry Van)
EQUIPMENTGROUP	Categorical group of truck types (Van)
STOPCOUNT	Number of stops for the shipment (min 2)
TOTALCUSTOMERRATE	The total amount NTG charges its customer
TOTALCARRIERRATE	The total amount paid to the carrier
PRODUCTCATEGORY	Grouping of different shipment types (Plastics et al.)
CARRIERID	Internal unique identifier for booked carrier
CUSTOMERID	Internal unique identifier for shipping customer
MILES	Total distance of shipment
FLASH_OFFER	Internal rate offered from NTG to the carrier through Beon Carrier platform
FLASH_OFFER_DATE	Date time of flash offer
BID_SYSTEM	Denotes if carrier engaged on either Beon mobile app or online portal
BIDS	Number of bids placed by carriers for a shipment
LOWEST_BID	Lowest bid value (in dollars)
LOWEST_BID_DATE	Date time of lowest bid
LOWEST_BID_SOURCE	Source of lowest bid (internal or Beon)
SECOND_LOWEST_BID	2nd lowest bid
SECOND_LOWEST_BID_DATE	Date time of second lowest bid
SECOND_LOWEST_BID _SOURCE	Source of second lowest bid (internal or Beon)
BOOKED_ON_BEON	True or False value if the shipment was booked on the Beon platform

3.2 Data Cleaning & Processing

The dataset was delivered in Excel. To properly model the data using machine learning, I needed to upload it into a Python instance and begin cleaning and converting the information.

Though the dataset was narrowed from the original 2 million records, the analysis period still contained 390,000 records. However, not every transaction was eligible for the digital booking platform. An extensive list of criteria must be met for a shipment to appear to eligible carriers on the platform, which led to further parsing of the data set. Other steps including removing typos and inconsistencies in location information (i.e., capitalizing some state abbreviations) were also taken.

Data Cleaning Process Steps:

1. **Data Import:** NTG provided data on each freight transaction since the introduction of their digital platform three years ago. The first step was to import the data from the years 2020 to 2023, using the Python library pandas. The data set included every freight transaction within the time frame, encompassing over 2 million transactions.
2. **Date Frame Selection:** Based on the platform relaunch insight from NTG, the targeted date frame spanned the most recent 6 months of transaction data from May 1, 2023, to Oct 31, 2023. The data frame was filtered to this range.
3. **Column Selection:** The next step was to select relevant columns from the dataset, focusing on essential features like load ID, pickup and delivery details, rates, equipment type, and carrier information. In this process, I found most columns had a fair chance of being relevant, and the data set was narrowed to 38 columns.
4. **Identify Feature:** 'Booked On Beon' is the target feature and was converted into a binary variable. This would allow for classification models to easily identify and operate with the target.
5. **Data Formatting:** The datasets underwent formatting adjustments for better utilization. This included converting dates to the datetime format for easy manipulation, categorizing certain textual data for efficient processing, and changing some columns to Boolean, integer, and float data types as appropriate.
6. **Data Exclusion:** Specific criteria were applied to exclude data that did not meet the project's requirements. Some of these examples include:
    a. Eliminate any shipment where the booked carrier of record was NTG
    b. Eliminate shipping modes that are not available on Beon Carrier
        i. Drayage, Intermodal & Rail, Storage, and Other

    c. Eliminate shipments with milage of 0. These shipments typically represent some type of outlier and not a standard point-to-point movement.

7. **Create Engineered Variables:** Expanding on the dataset, I created four new features that I believed had a chance to improve the prediction power of the model.
    a. Active To Booked: measures in seconds the time between the load's first activation and the load booking. This a key measure for how long the shipment was available for carriers to book the loads
    b. Rate per Mile: the total carrier rate divided by the total mileage of the shipment
    c. Booked to Pickup: measuring the time between when the load was booked and the scheduled pickup time. This gives an understanding of the lead time a carrier had when making their choice
    d. Margin: the total customer rate less the total carrier rate. While this is only an internal measure, it can frequently signal the desirability of a shipment.

After these steps, the data set to be analyzed contained 68,000 records. With this sample, I then began preparing the data for model creation. Additionally, after an initial round of analysis and understanding the high importance of the role played by Active to Booked in the model, I created another engineered feature: Active To Booked Bins. The purpose of this was to group the Active To Booked data into four bins to help improve the model classification. These bins, based on my experience and intuition working in brokerage for over a decade, were 0-30 minutes, 30-60 minutes, 60-180 minutes, and>180 minutes.

3.3 Model Preparation

With the data cleaned, organized, and prepared, the initial data exploration process was ready to begin. Data exploration provides the first step towards creating a functional model. In order to effectively prepare the data for the machine learning models, I created a preprocessing pipeline. This pipeline allowed for simple application of the data set into each ML model without the need for additional cleaning and processing for each model.

Preprocessing Steps:

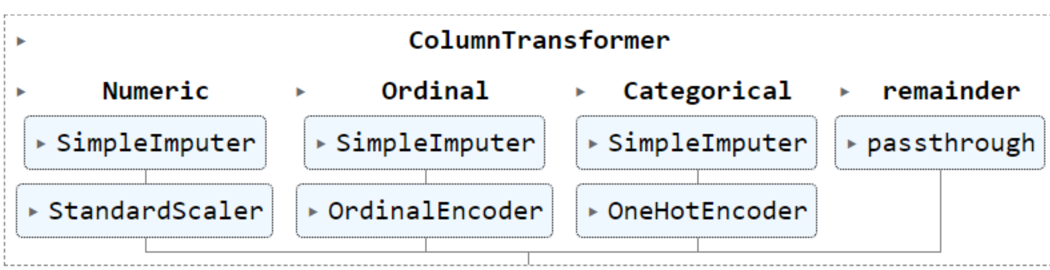
1. Separate features into three categories
    a. Numerical values: These are the features that consist of number values (i.e., Miles)

   b. Categorical values: These are the features that consist of categorical values with strings of characters (i.e., Equipment Group)
   c. Ordinal values: This is similar to categorical values, except there is an inherent ranking to these categories (i.e., Active to Booked Bins)
2. The Numerical values were then processed with
   a. Simple Imputer: Filled missing values with a median value
   b. Standard Scaler: Standardizes the numerical features by removing the mean and scaling to unit variance, which helps certain algorithms converge faster.
3. The Ordinal values were then processed with
   a. Simple Imputer: Filled missing values with the most frequent value
   b. Ordinal Encoder: converted the categorical feature to an ordinal integer value. This is useful when the categorical variable represents some order or rank.
4. The Categorical values were then processed with
   a. Simple Imputer: Filled missing values with the most frequent value
   b. One Hot Encoder: transformed the feature into a binary matrix, creating a new binary column for each category in the original data.
5. The remaining values were then passed through

Figure 1 shows a schematic of the data preprocessing pipeline used for the machine learning models used in this research and testing.

Figure 1

Model Preprocessing Pipeline



3.4 Machine Learning Applications and Approaches

The methodology for this project incorporates various Python-based ML tools, each chosen for its specific strengths. Gradient Boosting builds predictors in a stagewise manner, focusing on errors made by previous predictors and improving upon them. This approach makes it the preferred method for decision stumps. XGBoost stands out for its effectiveness in handling large datasets. It simplifies the prediction of weights in gradient boosting, making it more accessible and effective for extensive data scenarios like ours. Decision Trees offer a structured approach to operational decisions, useful in scenarios like carrier selection. Finally, Random Forests are known for their accuracy in complex prediction tasks. They have shown success in demand forecasting in the dynamic freight market and can provide valuable insight into the objectives of this research (Forsyth, 2019).

3.4.1 Logistic Regression

Logistic regression is a widely used supervised learning algorithm for classification and predictive analytics. It estimates the probability of an event occurring based on a given dataset of independent variables. The outcome is a probability, and the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied to the odds (the likelihood of success divided by the probability of failure). The model is commonly estimated via maximum likelihood estimation (MLE), which optimizes the best fit of log odds. Once the optimal coefficients are found, conditional probabilities for each observation can be calculated, yielding predicted probabilities. For binary classification, a probability less than 0.5 predicts class 0, while a probability greater than 0.5 predicts class 1 (What Is Logistic Regression? | IBM, 2024).

Logistic regression is particularly useful when dealing with categorical outcomes. with categorical outcomes, such as predicting whether a shipment will be booked digitally or not in the logistics domain.

3.4.2 Random Forest

Random Forest is a powerful ensemble learning algorithm that combines multiple decision trees to make predictions. It handles both classification and regression tasks. Each decision tree in the forest provides its prediction, and the final prediction is determined by aggregating the results from all trees. Random Forests are effective because they reduce overfitting and improve prediction accuracy. They utilize bagging (bootstrap aggregation) and feature randomness to create an uncorrelated forest of decision trees. Feature randomness ensures low correlation among trees, making Random Forests robust and accurate (What Is Random Forest? | IBM, 2024). These features make Random Forests an excellent choice for the type of classification problem being evaluated with this research.

3.4.3 Random Forest

K-Nearest Neighbors (KNN) is a straightforward yet potent method used in machine learning for classification and regression tasks. It operates on the principle that similar items are often close to one another in data space. In KNN, a target pattern's label is determined based on the labels of its K-nearest neighbors, with 'K' representing the number of neighbors considered. This count of neighbors influences the model's sensitivity: a small K value can make the model sensitive to noise, while a large K value can smooth over details by considering a broader swath of neighboring points. KNN involves measuring the distance between points, typically using metrics like Euclidean distance. This simplicity, coupled with its effectiveness in handling large datasets with low-dimensional feature spaces, makes KNN a widely utilized technique across various applications in pattern recognition and machine learning (Kramer, 2013)

KNN models the relationship between shipment characteristics and carrier preferences based on historical data. By identifying patterns in what carriers prefer, KNN can help predict the likelihood of new shipments being booked digitally.

3.4.4 Gradient Boosting with XGBoost

A more advanced model for machine learning with decision trees is called XGBoost. An abbreviation for eXtreme Gradient Boosting, XGBoost represents a sophisticated machine learning algorithm acclaimed for its effectiveness in processing structured and tabular datasets. Developed by Chen and Guestrin (2016), XGBoost is an advanced implementation of gradient-boosted decision trees, which is distinguished by its optimization for speed and performance. The algorithm's mathematical foundation incorporates several pivotal elements, including the objective function, gradient boosting methodology, and regularization techniques, each contributing to its robust predictive capabilities.

Gradient boosting, a cornerstone of XGBoost, iteratively refines predictions by consecutively incorporating new trees seen in Figure 3 designed to predict the preceding model's residuals or gradients (Friedman, 2001).

Furthermore, XGBoost integrates advanced regularization mechanisms, fundamentally contributing to its exemplary performance by curtailing the risk of overfitting. The regularization term imposes penalties on the number of tree leaves and the magnitude of leaf scores, thereby enforcing model simplicity and robustness (Chen & Guestrin, 2016).

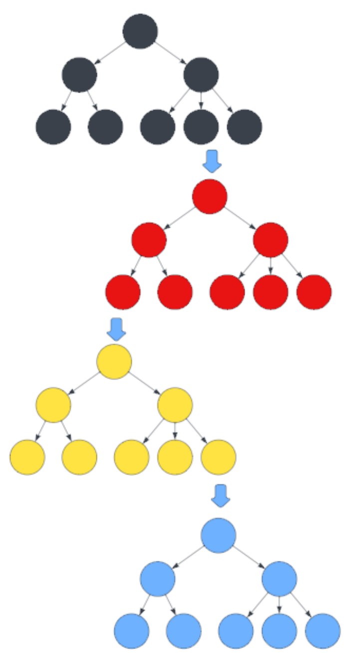
XGBoost is a powerful, scalable machine learning algorithm that provides high performance and speed in handling large datasets. Its ability to handle a mix of categorical and numerical features makes it well-suited for analyzing shipment characteristics. This aligns perfectly with the objectives to understand and predict carrier preferences on digital platforms. These properties make XGBoost the model of choice for this research

To apply gradient boosting with XGBoost to the data set, I followed these initial steps:

1. **Feature Selection:** Using my real-world experience, I selected an initial group of significant features that influence carrier booking decisions, such as stop count, bids, equipment type, and carrier rates.
2. **Data Encoding:** Using the preprocessing pipeline, the data was encoded and prepared for the model.
3. **Dataset Splitting**: The data was split into training and testing sets. This partitioning is vital for training the model on one set of data and validating its performance on another, ensuring the model's generalizability.
4. **Model Training and Prediction**: The XGBoost model was trained on the training set. This model is known for its efficiency in handling large-scale data and its ability to perform gradient boosting, which is pivotal for the predictive analysis.
5. **Performance Evaluation:** The model's performance was assessed on the test data, focusing on accuracy and other relevant metrics to ensure its efficacy in real-world scenarios.

Figure 3

Illustration of XGBoost Iterations



Note. Each iterative tree in the sequence builds upon the previous tree's performance, gradually improving performance.

3.5 Model Evaluation and Refinement

The evaluation phase was crucial in ensuring that each model was not only accurate but also reliable and applicable to the real-world complexities of freight brokerage. This phase involved rigorous testing in diverse scenarios and continuous refinement to adapt to the evolving nature of the industry.

The XGBoost model initially provided the most promising results. This being the case, further refining and parameter tuning was done to further improve model performance.

1. **Feature Refinement:** Selecting the right combination of features was crucial for the model's performance. First, based on a combination of personal experience and trial-and-error, different features were selected and tested. Each set was recorded along with model performance. This process improved the model's prediction scores, narrowing down the features to 13 key features.
2. **Further Feature Refinement:** After pursuing several iterations of guess-and-check feature selection, I took a more systematic approach to feature refinement. Using Python, I modeled every combination of the 13 features and measured model performance and accuracy. The process executed 8,191 iterations of the model and determined that the best performance came when all 13 of the previously expert-identified key features were included.
3. **Balancing Feature Selection:** While there was a temptation to include more features to improve the model's accuracy, this can lead to overfitting, where the model performs well on training data but poorly on unseen data. To avoid overfitting, I balanced the inclusion of informative features with the model's ability to generalize. Importantly, I avoided any features that have might undermined the model's credibility by accidentally 'giving away' the correct answers.
4. **Hyper Parameter Tuning:** To further improve model performance, I finely tuned each parameter of the model.
5. **Max Depth:** The first parameter I tuned was the Max Depth of the model. This parameter regulates the maximum tree depth and helps to balance accuracy with overfitting. The results can be seen in Figure 4.
6. **N Estimators:** The next tuning parameter was the n estimators. This feature regulates the number of boosting iterations the model performs. The results can be seen in Figure 5.
7. **Cross-Validation:** Cross-validation was used to assess the generalizability of the model. Training and testing the model on different subsets of the data ensured that the model's performance was consistent across various data samples. Cross-validation was applied to the model to evaluate performance.

8. **Iterative Refinement:** The model was continuously refined by experimenting with different combinations of features, assessing the model's performance, and adjusting accordingly. This iterative process was done to achieve the optimal balance between accuracy and generalizability.

SHAP values were then applied and evaluated to understand the importance of performance and features. Based on cooperative game theory, SHAP values provide a way to explain the output of machine learning models. They quantify the contribution of each feature to the prediction of each instance. In the context of this project, SHAP values reveal how different factors such as stop count, bids, or carrier rates individually impact the model's prediction of load acceptance rates (Lundberg & Lee, 2017).

After training the XGBoost model, SHAP values were calculated to interpret the model's decisions. This involved analyzing which features predominantly influence the model towards or away from predicting a specific outcome. In this case that outcome was whether or not a carrier books a load digitally. The use of SHAP values allowed for a deeper understanding of the model's predictive behavior, ensuring transparency and trust in its outputs. It aided in providing insights into influential factors, thereby guiding us in understanding the model's predictions more clearly.

With the model selection and interpretation phases complete, the next step was to review and evaluate the results.

Figure 4

Model Accuracy vs. Max Depth

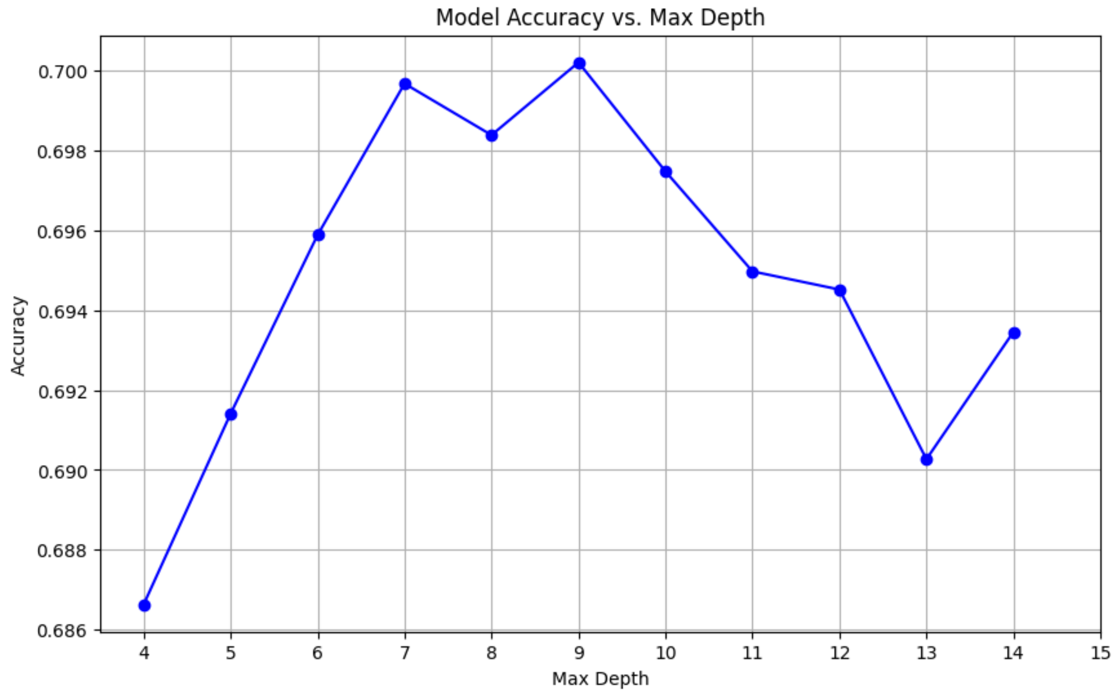
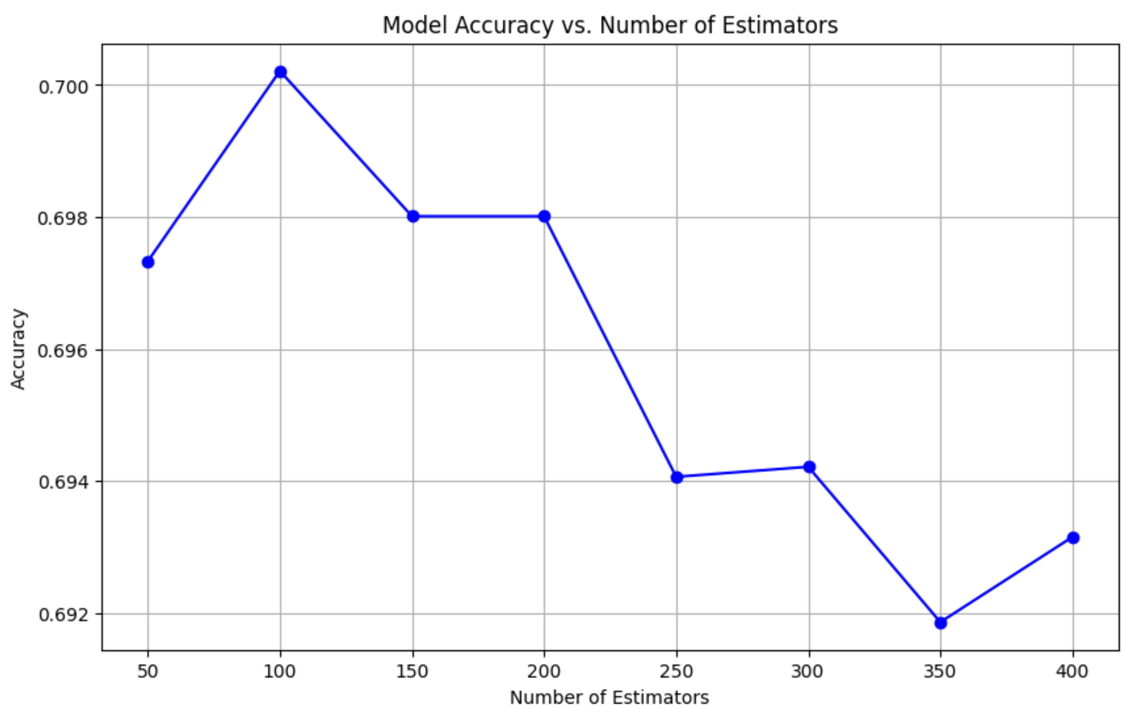


Figure 5

Model Accuracy vs Number of Estimators



4. RESULTS

The models performed well, producing compelling results. The strongest performance came from the XGBoost Classifier, which outperformed the Random Forrest, Logistic Regression and K-nearest neighbors.

4.1 XGBoost Model

The XGBoost Classifier model showed the highest accuracy results of the models tested in this research. The model showed good accuracy and prediction power. Table 2 shows the classification report for the XGBoost Classifier.

Table 2

XGBoost Classifier Results

	XGBoost Classifier			
Class	**Precision** %	**Recall** %	**f_1-Score** %	**Support**
False	67	48	56	5,239
True	71	85	77	7,937
Accuracy			70	13,176
Macro Avg	69	66	67	13,176
Weighted Avg	70	70	69	13,176

4.1.1 Receiver Operating Characteristic Curve

The ROC curve has an AUC (Area Under the Curve) of 0.74, which suggests that the model is effective at distinguishing between the positive class and the negative class.

The true positive rate (TPR), or recall or sensitivity, is plotted on the Y-axis, and the false positive rate (FPR) is plotted on the X-axis. The TPR indicates the proportion of actual positives that the model correctly identifies, while the FPR is the proportion of actual negatives that are incorrectly identified as positive. (*ADD CITE)

The ROC curve above the dashed line in Figure 6 represents a completely random classifier (AUC = 0.5). An ROC curve closer to the top-left corner indicates a better performance, where the TPR is high and the FPR is low. This curve shows that the model performs significantly better than random guessing but is not perfect. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. Conversely, the closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

4.1.2 Confusion Matrix

The confusion matrix in Figure 7 shows the performance of a classification model. Each component of the squared is as follows:

- Top-Left Square (True Negative): The model predicted “No” accurately 2,510 times.
- Top-Right Square (False Positive): The model predicted “Yes” inaccurately 2,729 times.
- Bottom-Left Square (False Negative): The model predicted “No” inaccurately 1,221
- Bottom-Right Square (True Positive): The model predicted “Yes” accurately 6,716

Figure 6

ROC Curve for XGBoost Model

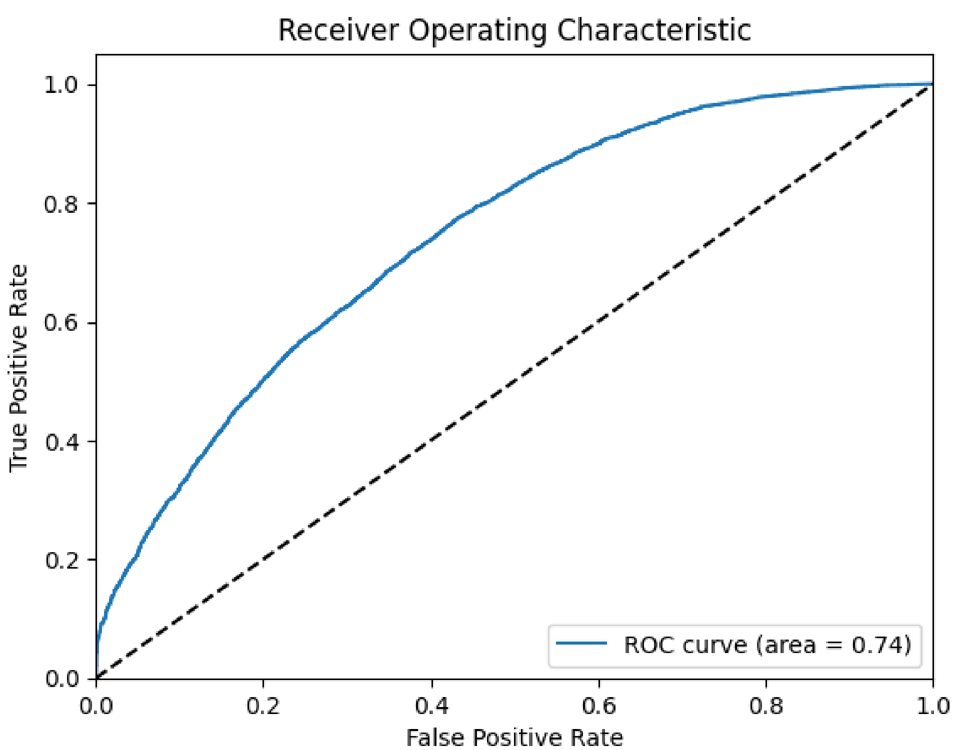
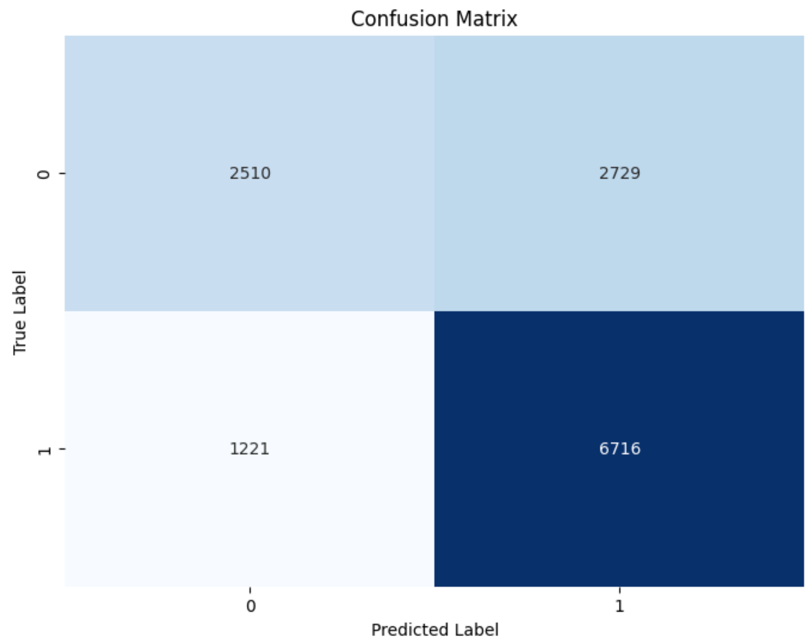


Figure 7

Confusion Matrix



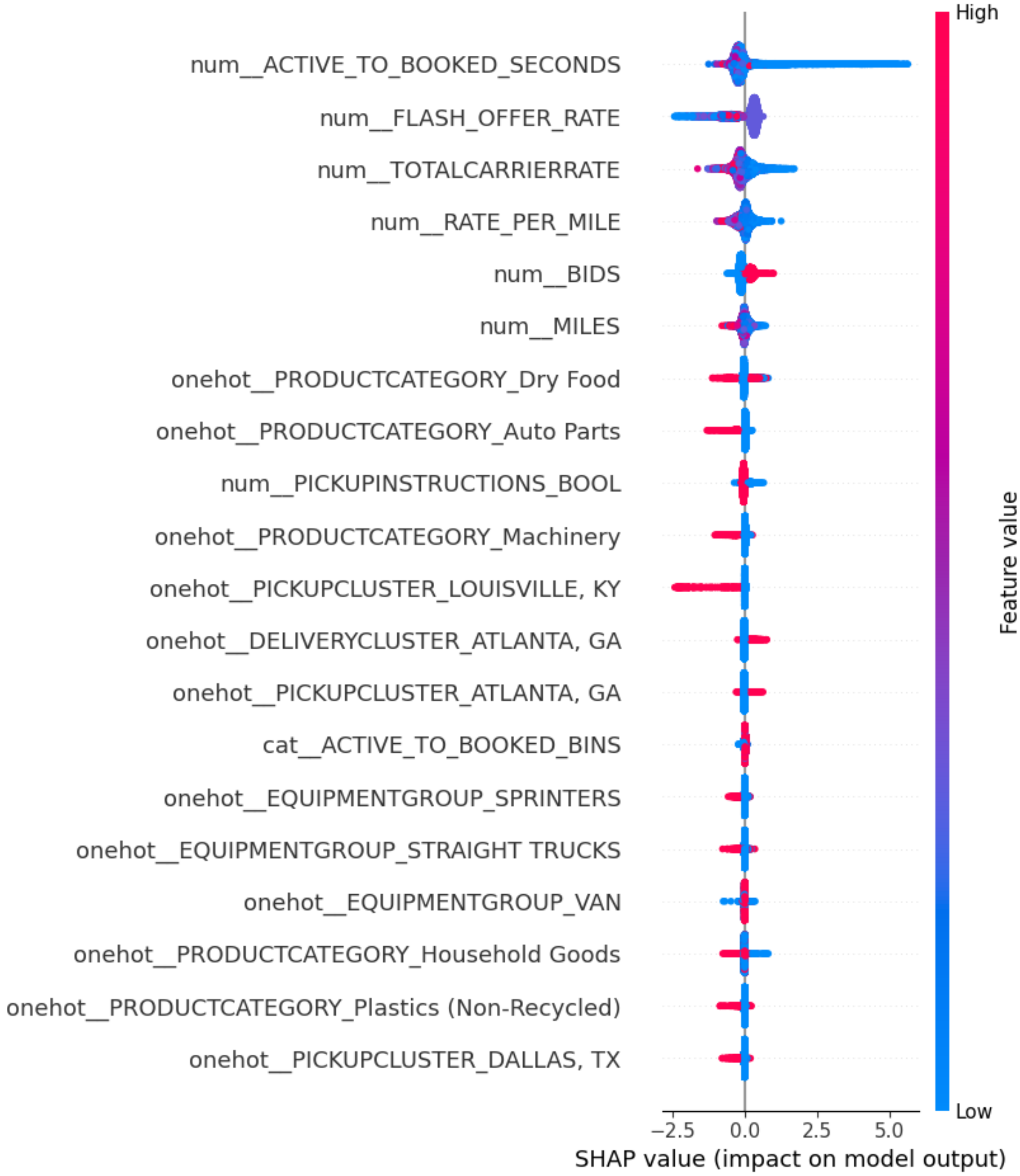
4.1.3 SHAP Values

The XGBoost classifier identified the features with the highest contribution to the model's predictions.

- Active to Booked: Low values for this feature had a high level of importance in the model predicting positive outcomes. While not all low values had this impact, it is important to note that no high values did. So, a low Active to Booked time means a high likelihood of being booked digitally.
- Flash Offer Rate: Low Flash Offer Rate values had a high impact on negative predictions in the model. Simply put, low dollar rates counter-offered to carriers negatively impacted their bookings on the digital platform.
- Total Carrier Rate: Low values of Total Carrier Rate were highly important for positive booking outcomes in the model. While this does not intuitively make sense, I will discuss some likely causes in later sections.
- Bids: A high number of bids had a high positive impact on the model, and a low number of bids had a high negative impact. This feature shows the most explicit divide between high and low and positive and negative outcomes.

Figure 8

SHAP Values of Feature Importance



Noteworthy in this SHAP diagram is the ACTIVE_TO_BOOKED_SECONDS. The low value of this feature indicates an outsized high impact on the model output. To put it simply, shipments booked digitally have a low booking time. Therefore, these shipments are booked quickly. This is a new and novel insight.

4.2 Other Models

Overall, the XGBoost model performed the best, which warranted a deeper analytic dive. I ran several other ML models to test their performance on the data set. The results of these tests for each model are described below.

4.2.1 Random Forest Classifier

The Random Forest Classifier model showed slightly less accuracy than the XGBoost model, with an accuracy store 2 points lower at 68%. It also showed lower results for precision and recall of both true and false cases. The classification report in Table 3 shows these results.

4.2.2 Logistic Regression

The logistic regression continued the trend of reduced accuracy of outpoint compared to previous models. This model showed a significant drop in accuracy compared to the Random Forest Classifier, with four 4-point lower accuracy. It was also significantly lower than the XGBoost model, with a 6-point lower accuracy. The logistic regression also had a significant drop in the False Recall %, falling to 29%. The classification report in Table 4 shows these results.

4.2.3 K-Nearest Neighbors

The classification report for K-Nearest Neighbors showed it to be tied for the least accurate of the models tested. The accuracy score of 64% in Table 5 was tied with the score of the Logistic Regression Model. There was an improvement compared to the logistic regression model in the recall of false values. However, those gains were traded for the Recall of True values.

With the objective of finding the model with the best predictive accuracy, the accuracy score of each model is the most important metric to compare. While each of the other models evaluated provided function results and valuable insight, they failed to perform at the level of the XGBoost model. This performance validates XGBoost as the model of choice for this problem.

Table 3

Random Forest Classifier Results

	Random Forest Classifier			
Class	**Precision** %	**Recall** %	**f_1-Score** %	**Support**
False	64	48	55	5,239
True	71	82	76	7,937
Accuracy			68	13,176
Macro Avg	67	65	65	13,176
Weighted Avg	68	68	67	13,176

Table 4

Logistic Regression Model Results

	Logistic Regression Model			
Class	**Precision** %	**Recall** %	**f_1-Score** %	**Support**
False	60	29	39	5,239
True	65	87	75	7,937
Accuracy			64	13,176
Macro Avg	63	58	57	13,176
Weighted Avg	63	64	61	13,176

Table 5

k-Nearest Neighbors Classifier Results

	kNN Classifier			
Class	**Precision** %	**Recall** %	**f_1-Score** %	**Support**
False	55	45	50	5,239
True	68	76	72	7,937
Accuracy			64	13,176
Macro Avg	61	60	61	13,176
Weighted Avg	63	64	63	13,176

5. DISCUSSION

5.1 Significance of Findings

The XGBoost model provides a functional first step for a proof-of-concept model that can continue to be refined and potentially productionalized. This model gives above-average prediction capability to predict whether an individual shipment will be booked digitally or not.
What is more compelling is a novel insight: digital shipments are booked quickly.

This pivotal discovery from this analysis with Nolan Transportation Group (NTG) revolves around the significant influence of the 'Time to Book' on the likelihood of digital booking. Through the analysis of the XGBoost model, the feature importance of ACTIVE_TO_BOOKED time is significantly high when this time frame has a low value. In practical terms, digital shipments get booked quickly.

This finding emerges as a linchpin in understanding digital freight brokerage dynamics. Specifically, my analysis elucidates that shipments designated for digital booking tend to be secured more rapidly than those handled through traditional methods. This insight has substantial implications for operational strategy within digital freight platforms.

Central to my findings is the observation that the probability of a shipment being booked digitally diminishes markedly as time progresses. This fact underscores a critical operational insight: there exists a relatively narrow window post-listing during which digital bookings are most likely to occur. After this period, the likelihood of digital engagement drops significantly, indicating that prompt action is crucial to capitalize on digital booking opportunities.

A mature freight brokerage typically has internal competition for booking spot shipments. Carrier Sales brokers, working on commission, react quickly to spot shipments to ensure their carrier is awarded the load. However, when brokerage leaders are armed with this new insight about digital booking, it becomes clear that the internal competition can stunt the growth of a brokerage's digital platform if this early booking window is not protected.

5.2 Recommendations for Nolan Transportation Group

Based on these findings, it is recommended that NTG focus on strategies that amplify the efficacy of their digital booking process. The strategies can benefit any brokerage looking to improve the adoption of their digital platform.

1. **Protect Early Activation Period:** The findings show how critical the initial period of shipment availability is towards digital booking success. Creating guardrails to ensure internal load

booking agents cannot access these shipments during an initial period will enhance digital adoption without the risk of reductions in service.

2. **Display Insights Directly on Digital Platform:** Shipments with a high probability of digital booking appear to be more desirable freight. By displaying information about the expected demand for a shipment externally on the digital platform, the brokerage can ensure interested carriers understand that the shipment they are considering may last only a short time. This information will further create a sense of urgency and improve digital booking success.
3. **Enhance Dynamic Pricing Models:** With the insight that digital shipments tend to move quickly immediately following their first availability, pricing models can be updated to take advantage of this behavior. Incorporating an initial but rapidly expiring discount to carrier offer rates can provide financial gain. Carriers who are acting quickly to secure digital shipments will have less ability to be price sensitive. With a rapid expiration of this discount, the brokerage can again ensure no loss of capacity or service.

It will be important to develop and refine the model to further improve the digital booking process. Once these operational recommendations are implemented, continued monitoring and care will be needed. Improving these tools and processes will provide both brokers and carriers with a more dynamic and responsive booking environment, optimizing the chances of a shipment being digitally booked.

5.3 Limitations

While the insights garnered from this study are compelling, several limitations must be acknowledged. Firstly, the research focused exclusively on the booking data of Nolan Transportation Group (NTG), which may restrict the generalizability of the findings. Although the digital booking trends identified here are likely applicable to similar logistics companies, idiosyncratic features specific to NTG could skew the results. These unique characteristics include internal company policies, specific client demographics, or peculiarities in the regional markets they primarily serve.

Additionally, as previously noted, the predictive model's accuracy is limited by the scope and quality of the data provided. External market factors such as economic fluctuations or sudden changes in supply chain demands, which could significantly impact booking behaviors, still need to be fully captured in the dataset.

5.4 Future Research

Looking forward, several avenues for further research present themselves. Expanding the study to include multiple logistics firms would help validate the generalizability of the findings and highlight firm-specific booking behaviors. This broader dataset could provide a more comprehensive view of the digital booking landscape across the freight brokerage industry.

Another promising area for future research lies in developing predictive models that more accurately forecast the time it takes for a shipment to get booked. Enhancing the prediction of booking time frames can provide significant operational benefits, such as improving resource allocation, optimizing carrier schedules, and reducing the time shipments spend idle. Such research could employ advanced machine learning techniques and incorporate real-time data streams to adjust predictions dynamically in response to changing market conditions.

Moreover, investigating the interaction effects between 'Time to Book' and other features such as shipment size, destination proximity, and carrier availability would refine the predictive accuracy further. Delving into these interactions can yield more profound insights into the complex dynamics that influence digital booking processes, offering more targeted strategies for improving digital engagement rates within the freight brokerage sector.

Overall, while this study offers foundational insights into the dynamics of digital booking, there remains substantial room to expand this research to harness deeper, more actionable insights. These efforts will enhance the theoretical understanding of digital booking dynamics and provide practical tools and strategies to optimize digital transactions in the brokerage industry.

6. CONCLUSION

This capstone project has delivered insight into the evolving dynamics within the freight brokerage industry with the help of machine learning. As the industry witnesses a shift from traditional brokerage methods to digital platforms, this study not only uncovers novel insights influencing carrier preferences, but also provides a useful tool to predict digital booking behaviors.

The freight industry is at a crossroads, driven by technological advancements that redefine operational paradigms. NTG's Beon Carrier and similar digital platforms are revolutionizing how carriers engage with freight opportunities, emphasizing speed and transparency. The study leveraged nearly 2 million transaction records to develop a machine learning model, revealing a critical insight: the probability of digital bookings decreases as the time a load remains unbooked increases, indicating a crucial, brief window for digital engagements.

Advanced machine learning techniques, especially Gradient Boosting with XGBoost, have yielded profound insights into the factors that significantly influence digital booking decisions. These insights are instrumental in shaping strategies for timing bookings, pricing shipments, and meeting specific shipment requirements. This knowledge provides actionable guidance for NTG and others in the industry to refine their digital services and boost operational efficiency.

The implications of this research extend beyond NTG, offering valuable lessons for the broader freight brokerage industry and underscore the importance of understanding carrier preferences in an increasingly digital landscape. The findings advocate for protecting the initial booking period to optimize digital bookings, a strategy that could redefine industry practices and enhance the efficacy and alignment of digital platforms with carrier preferences. This research not only enriches the industry's knowledge base but also equips other firms to advance their digital transformation strategies.

In conclusion, this capstone project has effectively bridged a crucial knowledge gap in the freight brokerage industry, providing both theoretical insights and practical tools that leverage the power of machine learning to forecast and influence carrier booking preferences. As the industry continues to navigate the challenges and opportunities presented by digitalization, the findings from this research will undoubtedly serve as a helpful tool for future research and strategic decisions within the digital freight brokerage domain.

REFERENCES

Acocella, A., & Caplice, C. (2023). Research on truckload transportation procurement. Journal of Business Logistics.

Balakrishnan, A. (2017, May 18). Uber's new app for truckers could disrupt one of America's core industries. CNBC. https://www.cnbc.com/2017/05/18/uber-launches-uber-freight.html

BOSS Editorial. (n.d.). Focusing on customer experience and ensuring success for partners leads to unprecedented growth for Nolan Transportation Group. BOSS Magazine. Retrieved from https://thebossmagazine.com/nolan-transportation/

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). ACM, New York, NY, USA.

Forsyth, D. (2019). *Applied machine learning* (1st ed.). Springer Nature.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.

Granato, R. (2019, July 25). The Rise of Digital Freight Matching and How It Works. Supply Chain 24/7. https://www.supplychain247.com/article/the_rise_of_digital_freight_matching

Helguera Sánchez, I., & Hendra Mukti, L. P. (2018). Uberization effects on freight procurement. MIT Capstone Project.

Holland, F. (2022, April 21). Convoy, an on-demand trucking firm started by ex-Amazon executives, is now valued at $3.8 billion. CNBC. https://www.cnbc.com/2022/04/21/convoy-on-demand-trucking-firm-valued-at-38-billion.html

Kramer, O. (2013). K-Nearest Neighbors. In Dimensionality Reduction with Unsupervised Nearest Neighbors (Intelligent Systems Reference Library, Vol. 51). Springer, Berlin, Heidelberg.

Lundberg, S. M., & Lee, S.-I.. (2017). A unified approach to interpreting model predictions. 30.

What Is Logistic Regression? | IBM. (2024, April 22). What Is Logistic Regression? | IBM. https://www.ibm.com/topics/logistic-regression

What Is Random Forest?. (2024, April 2). *What Is Random Forest?*. https://www.ibm.com/topics/random-forest

Zhou, Y., & Wan, Y. (2021). Does the Sharing Economy Technology Disrupt Incumbents? Exploring the Influences of Mobile Digital Freight Matching Platforms on Road Freight Logistics Firms. Production and Operations Management.