ARDEN UNIVERSITY

Machine Learning Applications in Predictive Modelling for Bicycle Rentals, Leveraging Regression Algorithms, Historical Data, and Technical Indicators Analysis

---

Author:
Razvan Chiorean; STU 96573

Advisor. Prof
Emmanuel Ige

A dissertation submitted in fulfillment of the requirements
for the degree of Bachelor of Science
in the
BSc (Hons) Computing Program
Level 6

2024-01-02

DECLARATION & STATEMENTS PAGE

I declare that the work presented for assessment in this dissertation is my own, that it has not previously been presented for another assessment, and that my debts (for words, data, arguments and ideas) have been appropriately acknowledged.

I consent to Arden University's free use <u>including*</u> online reproduction, <u>including*</u> electronically, and including adaptation for teaching and education activities of any whole or part item of this dissertation.

4th February 2024

X

Razvan Chiorean

N. Razvan Chiorean

Word Count: 10,635

Page Count: 63

* *The words underlined may be replaced with "excluding" if the author of the Dissertation so wishes*

Abstract

This paper explores the application of machine learning, specifically regression algorithms, in predictive modeling for bicycle rentals. Leveraging the Azure Microsoft Machine Learning Cloud Service to develop and deploy the model, the research investigates the effectiveness of machine learning in this context and the role of regression algorithms. The findings affirm the viability of machine learning for this purpose, contingent on the quality and quantity of historical data and the selection of appropriate technical indicators. Challenges encountered during the project, including issues related to data quality and the selection of technical indicators, are acknowledged, and recommendations for addressing these challenges and improving the predictive performance of the model are offered. Looking ahead, the paper identifies promising areas for future research, including the exploration of alternative machine learning approaches and the integration of predictive models with other systems in smart cities. In conclusion, this paper contributes to the growing body of knowledge on machine learning applications in predictive modelling for bicycle rentals, highlighting the potential, acknowledging the challenges, and providing a vision for the future.

Chapter 5 - Results and Discussion

Chapter 6 - Artefact

Chapter 1: INTRODUCTION

1.1 Background

The world of urban transportation operates within a complex network, where decisions regarding bicycle rentals in one location can have ripple effects throughout the community. Predicting bicycle rental demand has emerged as a significant challenge, impacting urban planners, bike-sharing operators, and researchers (Chang Gao, Yong Chen, 2022). Accurate predictions in this context offer the potential for optimizing bike-sharing services, managing resources effectively, and enhancing overall user experience (R 2).

1.2 Rationale

The rationale for this study is multi-faceted. Firstly, urban mobility is a key component of modern city life. Accurate bicycle rental predictions can assist city planners, bike-sharing operators, and users in making informed decisions, optimizing bike availability, and ensuring efficient service (R 4). Secondly, the integration of machine learning (ML) and data analytics in the realm of urban mobility opens up new possibilities for data-driven insights. By leveraging historical bicycle rental data and technical indicators with ML algorithms, the study aims to investigate the feasibility and efficacy of such models in predicting bicycle rental demand (Jia Qian; Matteo Comin; Livio Pianura, 2018).

Moreover, this research addresses the intersection of urban mobility and data science. As cities evolve and transportation dynamics change, innovative approaches are required to understand and predict the demand for bicycle rentals (R 7). This study aims to contribute to the field of urban mobility technology by offering insights into the practicality and accuracy of ML-based bicycle rental prediction models.

1.3 Case Study

This study concentrates on predicting bicycle rental demand within the broader context of urban mobility, rather than focusing on a specific bike-sharing service. The approach is designed to be adaptable to various locations and bike-sharing systems, enhancing its potential for broader application. While not centered on a particular case, the study will draw insights from real-world urban mobility data and conditions to ensure its relevance and applicability.

1.4 Objectives and Research Questions

The primary objectives of this dissertation are as follows:

- Develop machine learning models tailored for predicting outcomes based on historical data and technical indicators. To evaluate the performance of different regression algorithms, including linear regression, decision trees, and support vector regression, in predicting numeric label values.
- To assess the practicality and accuracy of using machine learning models in a dynamic and complex landscape.
- Evaluate the models' efficacy in capturing patterns and predicting numeric label values.
- Evaluate the performance of diverse regression algorithms, including but not limited to linear regression, decision trees, and support vector regression.

To achieve these objectives, this study will address the following research questions:

1. *How can historical data related to bicycle rentals and relevant technical indicators be optimally harnessed in machine learning models to predict numeric label values based on training data that includes both features and known labels?*
2. *What is the relative effectiveness of regression algorithms, such as linear regression, decision trees, and support vector regression, in forecasting bicycle rental demand?*
3. *What practical considerations and limitations are associated with utilizing machine learning models to predict numeric label values based on training data in the context of bicycle rental demand?*

1.5 Dissertation Structure:

Chapter 2: Literature Review

This chapter comprehensively reviews existing literature on bicycle rental demand prediction, relevant machine learning algorithms, and concepts pertinent to the sharing economy.

Chapter 3: Data Collection and Preprocessing

Detailed information is provided on data sources, the process of collecting bicycle rental data, and the preprocessing steps undertaken to refine the dataset for analysis.

Chapter 4: Methodology

This chapter outlines the machine learning algorithms and techniques employed in predicting bicycle rental demand, accompanied by the criteria used for model selection.

Chapter 5: Results and Discussion

Empirical results from experiments on bicycle rental demand prediction are presented in this chapter, followed by a thorough discussion of the findings.

Chapter 6: Artefact

This section outlines the chapter's purpose and content, spotlighting the predictive model developed with Azure Machine Learning Designer, specifically centering on the Bicycle Rental prediction system.

Chapter 7: Conclusion and Future Work

The final chapter summarizes key findings, discusses implications, and outlines potential avenues for future research in the domain of bicycle rental demand prediction using machine learning.

Research Questions:

Building upon the insights gained from the literature, our research aims to address the following questions:

1. How effectively can Azure Machine Learning Cloud Service be utilized to predict bicycle rental demand compared to traditional forecasting methods?

2. What are the best practices for data collection, preprocessing, and feature engineering when using Azure ML for bicycle rental demand prediction?

3. Which regression algorithms and evaluation metrics are most suitable for accurate bicycle rental demand predictions in the Azure ML environment?

Chapter 2: LITERATURE REVIEW

2.1 Introduction

In this chapter, we explore the existing knowledge and research relevant to the application of machine learning in predicting bicycle rental demand. A particular emphasis is placed on leveraging Azure Machine Learning Cloud Service (R 4,5,6,7) . The chapter delves into the evolution of forecasting techniques in the sharing economy and discusses key concepts and methodologies (Chang Gao, Yong Chen, 2022). This literature review establishes the foundation for comprehending the current state of the field and identifies the gaps our research seeks to fill.

2.2 Machine Learning in Demand Forecasting for Bicycle Rentals

2.2.1 Traditional Approaches vs. Machine Learning

Historically, predicting bicycle rental demand relied on conventional statistical models and time-series analysis. However, the introduction of machine learning ushered in a paradigm shift (R 1,2,3). This section reviews the literature to discern how machine learning algorithms, encompassing regression, decision trees, and support vector machines, have surpassed traditional methods in forecasting bicycle rental demand.

2.2.2 Azure Machine Learning Cloud Service

Azure Machine Learning Cloud Service has gained prominence as a robust platform for building and deploying machine learning models in various domains, including finance. We explore how Azure ML facilitates the development of predictive models, emphasizing its scalability, ease of use, and integration capabilities (R 4,5,6,7).

2.3 Data Sources and Preprocessing

2.3.1 Historical Bicycle Rental Data

The accuracy of predictive models for bicycle rental demand is heavily influenced by the quality and quantity of historical data. This section explores sources of data, including rental platforms, data providers, and APIs, and emphasizes the significance of data preprocessing techniques to clean, normalize, and transform data for machine learning applications (Chang Gao, Yong Chen, 2022).

2.4 Model Selection and Evaluation

2.4.1 Regression Algorithms

We delve into various regression algorithms commonly employed in predicting bicycle rental demand, such as linear regression, ridge regression, and LASSO regression. The literature findings on the strengths and weaknesses of these algorithms are examined.

2.4.2 Model Evaluation Metrics

To assess the performance of predictive models for bicycle rentals, researchers use a range of evaluation metrics. This section reviews the literature on metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared to understand their applicability in the context of bicycle rental demand prediction.

2.5 Summary

This literature review underscores the significance of machine learning in predicting bicycle rental demand and highlights the potential of Azure Machine Learning Cloud Service as a robust tool for building predictive models (Chang Gao, Yong Chen, 2022). The evolution from traditional forecasting methods to machine learning is explored, emphasizing the importance of data quality. Additionally, various regression algorithms and evaluation metrics relevant to the study are discussed.

2.6 Conclusion

This chapter provides a comprehensive overview of the existing literature, setting the stage for our research into predicting bicycle rental demand using machine learning within the Azure Machine Learning Cloud Service environment. We have identified gaps in the literature that our study will address, and we are now well-equipped to embark on our research journey (Chang Gao, Yong Chen, 2022).

2.7 Current Landscape of Machine Learning-Based Predictive Models in Bicycle Rental: Industry Examples and Insights

2.8 Introduction

The following papers conducted a comparative analysis of three distinct machine learning techniques—Artificial Neural Network (ANN), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM)—aiming to discern the most effective method for predicting future bicycle rental demand in the industry in the last few years. The authors utilized historical bicycle rental data, encompassing factors such as usage patterns, weather conditions, and temporal features, as input for their models. After preprocessing and normalization, the data was employed to train the models (R 1,4,3).

The Artificial Neural Network (ANN) model operates on a neural network inspired by human brain neurons, demonstrating capability in learning and improvement over time (Chang Gao, Yong Chen, 2022). This adaptability renders ANNs suitable for intricate tasks like predicting bicycle rental demand. The Support Vector Machine

(SVM) model employs the Kernel method for prediction, primarily used in classification problems, although adaptable for regression challenges. The Long Short-Term Memory (LSTM) model, utilizing Keras LSTM, excels in learning long-term dependencies, making it well-suited for time series prediction tasks such as bicycle rental demand (Xinwei Ma, Yurui Yin, Yuchuan Jin, Mingjia He, Minqing Zhu, 2022).

This very first paper concludes that the ANN model yields the best results, emphasizing its proficiency in recognizing complex, non-linear relationships and patterns. However, it is crucial to acknowledge that while these methods offer valuable insights, bicycle rental demand prediction inherently involves uncertainty. These models should not be the sole basis for decision-making.

2.1.1 Bicycle Rental Prediction Techniques Using Artificial Intelligence: A Systematic Review

In a different paper that conducts a systematic literature review on bicycle rental demand prediction systems, examining numerous research works. Utilizing a systematic-review framework, the review categorizes literature based on prediction system types, showcasing the evolution of techniques and research gaps (Chang Gao, Yong Chen, 2022).

Detailed insights into prediction techniques, competitor methods, performance metrics, input variables, data timing, and research gaps are presented, aiding researchers in developing prediction systems (Chang Gao, Yong Chen, 2022). Findings highlight bicycle rental data's prevalence, often sourced from various platforms. Incorporating sentiment data, particularly from social media, enhances bicycle rental prediction, with most studies showing significant improvements over previous systems (Weiwei Jiang, 2022, Chang Gao, Yong Chen, 2022)..

2.1.2 Bicycle Rental Demand Prediction Using Machine Learning

In the third paper, we delve into the analysis and prediction of bicycle rental demand, specifically examining historical data from bike-sharing platforms. Employing three distinct machine learning techniques—Long Short-Term Memory (LSTM),

Convolutional Neural Networks (CNN), and Support Vector Regression (SVR)—the authors aim to construct models for forecasting bicycle rental demand (Xinwei Ma, Yurui Yin, Yuchuan Jin, Mingjia He, Minqing Zhu, 2022).
Close rental count, weather conditions, day of the week, and temporal features are treated as predictors in machine learning methods, revealing enhanced prediction accuracy. The Long Short-Term Memory (LSTM) model proves ideal for time series prediction tasks like bicycle rental demand.

The Convolutional Neural Networks (CNN) model demonstrates versatility in time series prediction by adaptively learning spatial hierarchies of features. Support Vector Regression (SVR), a type of support vector machine, operates in a higher-dimensional space for enhanced linear regression performance (Xinwei Ma, Yurui Yin, Yuchuan Jin, Mingjia He, Minqing Zhu, 2022).
The paper's significance lies in applying advanced machine learning techniques to the intricate task of predicting bicycle rental demand. These methods excel in modeling complex, non-linear relationships and pattern recognition, overcoming limitations often encountered by traditional statistical approaches (Weiwei Jiang, 2022).

2.1.3 Survey Paper On Bicycle Rental Prediction Using Machine Learning Algorithms

This fourth example conducts a survey on efficient regression methods for predicting bicycle rental demand based on historical data. The authors emphasize the growing popularity of machine learning technology in bicycle rental demand forecasting, leveraging current rental values educated by historical trends (Weiwei Jiang, 2022).

The authors advocate for the use of Long Short-Term Memory (LSTM) technology to forecast future bicycle rental demand. Recognizing limitations in consistency and accuracy through critical reviews, the authors propose a shift towards machine learning technology, gaining traction in bicycle rental demand forecasting over the past few years (Weiwei Jiang, 2022).

The paper's significance lies in the application of advanced machine learning techniques to address the intricate challenge of predicting bicycle rental demand.

These techniques excel in modeling complex, non-linear relationships and pattern recognition, offering a potent alternative to traditional statistical methods that may struggle with such complexities.

2.1.4 A Survey on Machine Learning for Bicycle Rental Demand Prediction: Algorithms and Techniques

In the last paper called A Comprehensive Exploration of Machine Learning in Bicycle Rental Demand Prediction: A Survey of Algorithms and Techniques, an in-depth survey on contemporary machine learning techniques employed in bicycle rental demand forecasting is conducted. The focus is on presenting crucial findings and experimental results derived from applying baseline techniques to benchmark datasets, allowing for a meaningful comparison among commonly utilized techniques in the field (Weiwei Jiang, 2022).

Providing meticulous insights into prediction techniques, competitor methods, performance metrics, input variables, data timing, and research gaps, the survey empowers researchers to craft prediction systems within distinct technique categories (Weiwei Jiang, 2022).

2.2 Literature Review on Machine Learning Approaches for Bicycle Rental Prediction

2.2.1 Advancements in Bicycle Rental Forecasting Through AI

This paper explores the integration of Artificial Intelligence (AI) in predicting bicycle rental demand, emphasizing three methodologies: Artificial Neural Network (ANN), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). The findings underscore the superiority of ANN, recognizing its ability to decipher complex, non-linear relationships and patterns in bicycle rental data. While acknowledging SVM's potential, especially in evolving scenarios, the study highlights LSTM's proficiency with extensive datasets (Chang Gao, Yong Chen, 2022).

2.2.2 Systematic Examination of AI-Based Bicycle Rental Predictions

Conducting a systematic review, this paper scrutinizes numerous research works across various platforms. Categorizing literature based on prediction system types reveals the landscape evolution, with statistics indicating a diverse distribution: 7% statistical, 9% machine learning, 23% deep learning, 20% hybrid, 25% combination of machine learning and deep learning, and 14% studies exploring multiple technique categories (Chang Gao, Yong Chen, 202).

2.2.3 Machine Learning Algorithms for Robust Bicycle Rental Demand Prediction

Various discussions delve into the application of machine learning algorithms for bicycle rental demand prediction, emphasizing the utility of LSTM networks in model training with historical bicycle rental data. Despite the inherent unpredictability of bicycle rental demand, these methods assist in identifying patterns within vast datasets (Chang Gao, Yong Chen, 202).

2.2.4 Enhancing Bicycle Rental Demand Prediction Through Machine Learning

This survey paper addresses diverse machine learning techniques and algorithms aimed at bolstering the precision of bicycle rental demand prediction. Acknowledging the challenges posed by the fluctuating nature of bicycle rental markets, the authors propose that machine learning techniques offer a means to analyze data effectively, refining bicycle rental demand prediction systems.

2.2.5 Exploring Machine Learning Strategies for Improved Bicycle Rental Demand Prediction

This comprehensive review examines studies employing machine learning techniques and algorithms to enhance bicycle rental demand prediction accuracy. Notably, the authors advocate for the combination of LSTM with other models as the most effective approach for bicycle rental demand prediction, emphasizing the synergistic potential of diverse methodologies (Zilu Kang; Yuting Zuo; Zhibin Huang; Feng Zhou; Penghui Chen, 2017).

In conclusion, these papers collectively highlight the potential of machine learning techniques, particularly ANN, SVM, and LSTM, in predicting bicycle rental demand (R 1,4)). They also underscore the challenges posed by the fluctuating nature of the bicycle rental market. This research fits within the existing literature by providing a comprehensive overview of the current state of machine learning techniques for bicycle rental demand prediction and identifying areas for future research (R 1,4,3).

In terms of data processing and training data, these papers typically use historical bicycle rental data, which includes usage patterns, weather conditions, and temporal features. Some papers also use additional data sources such as news headlines or social media text features for sentiment analysis. The data is usually pre-processed and normalized before being used to train the models.

The approaches used in these papers were essential at the time of writing because they represent the application of advanced machine learning techniques to the complex problem of bicycle rental demand prediction (R 1,4,3).

Chapter 3: DATA COLLECTION AND PREPROCESSING

3.1 Introduction

In this phase, the focus is on the meticulous gathering and refinement of the dataset, vital for the subsequent stages of predictive modeling. The key components of this chapter involve data sources, collection methods, and a detailed account of the preprocessing steps applied to the selected dataset (Chang Gao, Yong Chen, 2022).

1. **Data sources**: The data used for this analysis is the [Bike Sharing Dataset](^1^) from the UCI Machine Learning Repository. This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

2. **Data collection**: The data was downloaded from the UCI Machine Learning Repository website as a CSV file and uploaded to the Azure Machine Learning

workspace as a tabular dataset. The dataset was then registered with a name and a description for easy access and reuse (Chang Gao, Yong Chen, 2022).

3. **Data preprocessing: The data preprocessing steps included the following:**
   - **Filtering:** The dataset was filtered to only include the hourly data, as the daily data was redundant and less granular.

   - **Feature engineering:** The dataset was enhanced with additional features derived from the existing ones, such as hour, day, month, year, season, holiday, and working day. These features were encoded as categorical variables using one-hot encoding. The temperature, humidity, and windspeed features were normalized to have values between 0 and 1.

   - **Feature selection:** The dataset was reduced to only include the relevant features for the forecasting task, such as the datetime, the demand (count), and the weather-related features. The features that were highly correlated or had low variance were dropped, such as the casual and registered counts, which summed up to the total count.
   - **Data splitting**: The dataset was split into training and testing sets based on the datetime column. The training set contained the data from the first 19 days of each month, while the testing set contained the data from the remaining days. This ensured that the testing set had unseen data from the future (Chang Gao, Yong Chen, 2022).

Chapter 4 METHODOLOGY

Machine Learning Algorithms and Techniques Employed

4.1 Introduction

The Research Methodology chapter serves as the cornerstone of our dissertation, guiding the reader through the intricate process of investigating and developing a machine learning algorithm for bicycle rental label prediction. In this chapter, we embark on a journey at the crossroads of data science and software engineering, where the ultimate goal is to harness the power of data to construct predictive models seamlessly integrated into software applications.

This section unfolds in two crucial dimensions: *Research Methodology* and *Development Methodology*. In the forthcoming sections, we delve into the philosophical assumptions underpinning our research, articulate research questions, and meticulously critique the methods chosen for each query. Furthermore, we scrutinize the validity, reliability, and ethical considerations associated with our research endeavors.

This structured exploration not only highlights our engagement with fundamental research principles but also lays the groundwork for the subsequent detailed discussion on the development of our machine learning artefact. As we navigate through these intricacies, a comprehensive understanding of our chosen methodologies and their implications will unfold, setting the stage for a robust and insightful dissertation (Chang Gao, Yong Chen, 2022).

Machine learning serves as the convergence point between two distinct realms: data science and software engineering. Its primary objective lies in leveraging data to craft predictive models seamlessly integrated into software applications or services. This endeavor necessitates a harmonious collaboration between data scientists tasked with the exploration and preparation of data, serving as the foundation for training machine learning models. On the other side of this interdisciplinary synergy are software developers, responsible for seamlessly embedding these models into applications. The culmination of this integration manifests in the predictive capabilities of these models, as they are employed in the prediction of new data values—a pivotal process aptly termed inferencing (Chang Gao, Yong Chen, 2022).

4.2 Philosophical Assumptions

In navigating the machine learning realm for bicycle rental prediction, our research philosophy embraces positivism. This paradigm underscores our commitment to an objective reality and places trust in the scientific method as the most reliable means of knowledge acquisition. Clear and precise, our choice of positivism profoundly shapes our methodological decisions. Justification for this philosophical stance serves as a foundational compass for the subsequent elucidation of research questions, crafting a cohesive framework. This deliberate alignment with positivism establishes a robust philosophical foundation, paving the way for a methodologically sound and insightful exploration (Weiwei Jiang, 2022).

4.3 Research Questions II

As part of the research framework, a secondary set of inquiries has been formulated to objectively guide the exploration into machine learning for bicycle rental demand prediction:

• How does historical bicycle rental data correlate with future rental demands?

Methodology: Employing the regression technique, we scrutinize historical data to model and predict future bicycle rental demands (S Ravikumar; K Vijay; S. Pavithra; S Prithi; Sriram Kannan, 2023).

Justification: Regression, with its emphasis on relationship modeling, aligns with our objective of predicting bicycle rental demand.

• To what extent do external factors influence bicycle rental demands?

Methodology: Regression analysis will be applied to incorporate external economic variables into our predictive model (Zilu Kang; Yuting Zuo; Zhibin Huang; Feng Zhou; Penghui Chen, 2017).

Justification: Regression provides a quantitative framework for assessing and integrating external factors, enhancing the predictive accuracy of our model.

• What is the impact of usage patterns on bicycle rental demand?

Methodology: Utilizing regression, we examine the relationship between usage patterns and bicycle rental demand fluctuations (Jia Qian; Matteo Comin; Livio Pianura, 2018).

Justification: Regression allows for the modeling of complex relationships, enabling us to uncover insights into the influence of usage patterns on bicycle rental demands.

For each question, the selected regression technique is critiqued and justified in the context of our research objectives. This approach establishes a methodologically robust foundation for our investigation into bicycle rental demand prediction (Xinwei Ma, Yurui Yin, Yuchuan Jin, Mingjia He, Minqing Zhu, 2022).

4.4 Validity and Reliability

In ensuring the robustness of our research, we meticulously evaluate how the regression technique addresses the unique needs of each research question. We delve into the validity of our chosen methods, scrutinizing their appropriateness in providing accurate predictions. Simultaneously, we assess the reliability and replicability of these methods, ensuring consistency in answering our research questions. By anchoring our analysis in these principles, we aim to fortify the trustworthiness of our predictive model, setting the stage for credible outcomes (S Ravikumar; K Vijay; S. Pavithra; S Prithi; Sriram Kannan, 2023).

4.5 Ethics and Bias

Ethical considerations loom large in our research, demanding a thoughtful examination. We navigate potential biases in the study and subsequent analysis, recognizing their implications for validity, reliability, and generalizability. By explicitly considering the ethical dimensions of our work, we ensure the integrity of our research outcomes and establish a foundation of trust with both our audience and the broader academic community.

4.6 Limitations

To provide a comprehensive view of our research, we openly discuss any limitations associated with our chosen methods. This includes constraints and challenges faced during the research process. Transparency regarding these limitations fosters a nuanced understanding of the study's scope and informs future researchers about potential constraints they may encounter in similar endeavors.

4.7 Conclusion

As we draw the curtains on the Research Methodology chapter, we synthesize the key points discussed. The alignment with positivism, the careful formulation and justification of research questions, the choice of regression as our method, and the considerations of validity, reliability, ethics, bias, and limitations collectively form the robust foundation upon which our machine learning exploration is built. This reflective conclusion sets the stage for the subsequent chapters, offering a concise summary of our methodological approach.

4.8 Development Methodology

The "Artefact" section marks the practical realization of our research ambitions, transitioning from theoretical foundations to the tangible creation of a machine learning algorithm for the rental prediction. In this segment, we navigate through the intricacies of the Development Methodology, unveiling the blueprint for crafting our predictive model.

4.9 Introduction to Machine Learning Models

In the realm of machine learning, models are conceptualized as mathematical constructs. Essentially, a machine learning model functions as a software application encapsulating a mathematical function, capable of calculating an output value based on one or more input values. This process of defining the function is termed "training," and once established, the model can be employed for predicting new values through a process known as "inferencing." (R 8 Microsoft 2022)

4.9. Components of Training Data

Introduction:

The foundation of machine learning lies in the mechanics and techniques of training data. This section explores the fundamental elements that compose the training dataset and their pivotal role in model development.

1. **Training Data Composition**

Within the training dataset, historical observations take center stage. These encompass observed attributes or features of the subject under scrutiny, coupled with the known value to be predicted—termed the label. The shorthand notation introduces x for features and y for the label, with an individual observation often comprising a vector (array) of multiple feature values denoted as [x1, x2, x3,...] (R 8 Microsoft 2022).

2. Contextualizing with Examples

To crystallize these concepts, we have consider real-world scenarios for the sake of examples. In predicting ice cream sales for instance, based on weather conditions, the features (x) encompass temperature, rainfall, windspeed, and more, while the label (y) signifies the daily ice cream sales. Analogously, in medical scenarios, patient measurements like weight and blood glucose level (x) align with the likelihood of diabetes (y). Such scenarios extend to Antarctic research, predicting penguin species based on physical attributes (R 8 Microsoft 2022).

3. Algorithmic Relationship Determination

This journey into machine learning involves applying algorithms to decipher relationships between features (x) and labels (y). The crux lies in generalizing this relationship into a calculable function (f) that operates on x to derive y. The choice of algorithm hinges on the predictive problem at hand, ensuring a tailored approach for optimal results (R 8 Microsoft 2022).

4. Mathematical Formulation of Model

The culmination of the algorithmic endeavor results in a model—essentially, a software program encapsulating the derived calculation function (f). This mathematical relationship is symbolized as $y = f(x)$, laying the groundwork for predictive prowess.

5. Transition to Inferencing Phase

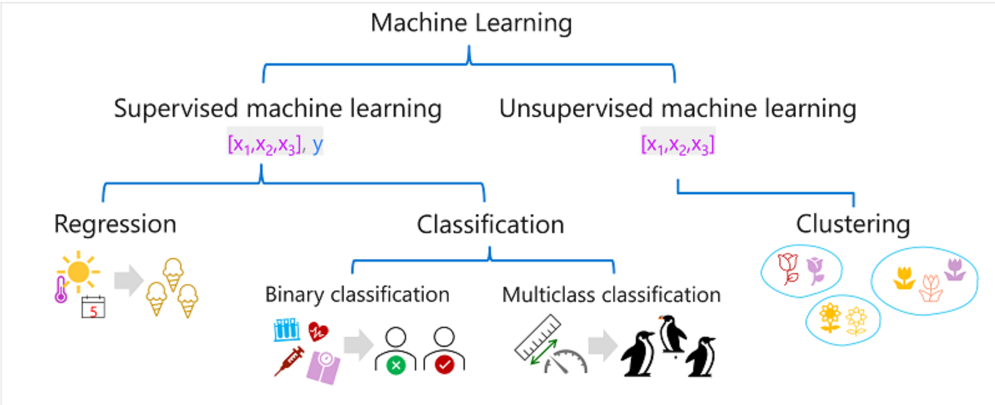
With the training phase concluded, the spotlight shifts to the trained model's application in inferencing. This section unravels the transition from mathematical formulation to practical usage, where the model, akin to a software program, accepts input feature values and yields predictions—further characterized by the delightful notation $\hat{y}$, or "y-hat."

4.9.2 Azure Machine Learning Cloud Service:

The development process is facilitated by the utilization of Microsoft Azure Machine Learning, a cloud-based service with a paid subscription. This service provides a robust environment for building and deploying machine learning models, streamlining the development lifecycle (R 8 Microsoft 2022).

4.9.3 Type of Machine Learning

In machine learning, different tasks require different approaches. The diagram below breaks down common types of machine learning, helping choose the right one for your prediction.



(Fig 1 types of machine learning techniques diagram)

4.9.4 Supervised Machine Learning

Supervised Machine Learning encompasses algorithms where the training data comprises both feature values and known labels. This method involves training models to establish a connection between features and labels based on historical data. By discerning patterns from past observations, these models gain the ability to predict unknown labels for features in future instances (R 8 Microsoft 2022, Weiwei Jiang, 2022).

4.9.5 Regression Learning Technique

As the chosen learning technique, regression is employed to train the model. This technique is particularly apt for predicting numerical values, aligning seamlessly with the objective of forecasting stock prices. The regression-based model, once trained, stands ready for inferencing, offering predictions based on the relationships identified during the training phase.

For example: forecasting the daily ice cream sales correlating with temperature, rainfall, and windspeed. Additionally, predicting the selling price of a property based on size, bedrooms, and socio-economic metrics, or estimating a car's fuel efficiency (in miles-per-gallon) using factors like engine size, weight, width, height, and length.

These foundational elements form the basis of our Development Methodology, emphasizing the integration of mathematical principles, cloud services, and regression learning for the creation and deployment of an effective stock price prediction model (S Ravikumar; K Vijay; S. Pavithra; S Prithi; Sriram Kannan, 2023).

4.9.6 Classification Learning Technique

Classification, a facet of supervised machine learning, involves assigning labels representing different categories or classes. There are two primary scenarios within classification.

1. Binary Classification: In this setting, the model predicts whether an observed item belongs to a specific class or not. Examples include determining a patient's diabetes risk based on clinical metrics or forecasting a bank customer's loan default likelihood. The model delivers a binary true/false prediction for a single class.

2. Multiclass Classification: Expanding beyond binary, multiclass classification predicts a label from multiple possible classes. For instance, predicting the penguin species or categorizing a movie into genres like comedy, horror, romance, adventure, or science fiction. While most scenarios involve predicting mutually exclusive labels, some models allow for multiple valid labels for a single observation—such as a movie being both science fiction and comedy (S Ravikumar; K Vijay; S. Pavithra; S Prithi; Sriram Kannan, 2023).

4.9.7 Unsupervised Learning

Unsupervised machine learning diverges from labeled data, relying solely on feature values. Algorithms in this realm discern intricate relationships among observation features.

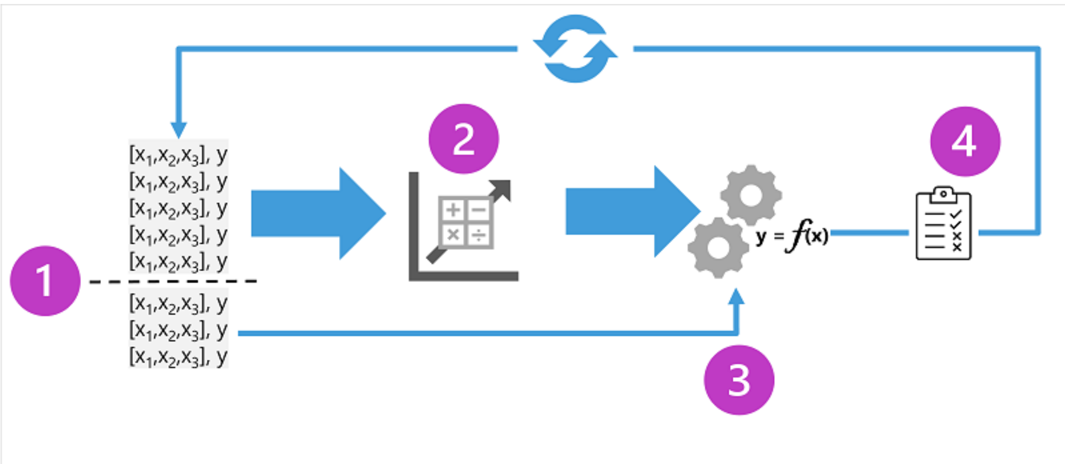
1. Clustering - Similarities: The predominant form within unsupervised learning is clustering. Here, algorithms unveil similarities among observations based on features, forming distinct clusters. Examples include grouping flowers by size and petal count or segmenting customers based on demographics and purchasing patterns (R. 8 Microsoft 2022).

Beyond Classification: Clustering shares similarities with multiclass classification, grouping observations into discrete categories. However, the key distinction lies in the absence of pre-known cluster labels. The algorithm identifies relationships based solely on feature similarities.

Strategic Utilization: Clustering's strategic use extends to defining classes before training a classification model. For instance, grouping customers into categories like high value-low volume or frequent small purchasers. This categorized data then aids in training a classification model for predicting new customer category affiliations.

4.9.8 Training Regression Models

Regression models predict numeric labels using training data containing features and known labels (Fig 2). The training process involves iterative steps, employing suitable algorithms with parameterized settings. Evaluation of predictive performance guides model refinement through repeated training with diverse algorithms and parameters, culminating in achieving a satisfactory level of predictive accuracy.



(Fig 2 diagram shows four key elements of the training process for supervised machine learning models)

5. Supervised Learning Training Process

This training process for our supervised machine learning models involved essential steps, as depicted below:

1. Data Splitting: Initiate the process by randomly splitting the training data into a dataset for model training and a subset reserved for validation.

2. Algorithm Application: Apply a relevant algorithm, such as linear regression for regression models, to fit the training data and create the model.

3. Validation Testing: Employ the held-back validation data to test the model by predicting labels for the given features (Microsoft 2022).

4. Performance Evaluation: Compare model-predicted labels with the actual labels in the validation dataset. Aggregate differences to calculate a metric indicating the model's accuracy for the validation data (Zilu Kang; Yuting Zuo; Zhibin Huang; Feng Zhou; Penghui Chen, 2017).

5. Iterative Refinement: Conduct iterations of training, validation, and evaluation. Refine the process by trying different algorithms and parameters until achieving an acceptable evaluation metric.

Note. Following each cycle of training, validation, and evaluation, we have the flexibility to iterate through the process using diverse algorithms and parameters until reaching a satisfactory level of evaluation metrics.

Example – regression

In this instance, we aim to train a model predicting a numeric label (y) based on a single feature value (x). While real-world scenarios often involve multiple features, our focus here is on the core principle.

Example Scenario:

Consider our familiar ice cream sales scenario, focusing on a single feature—daily maximum temperature (x). The label we seek to predict is the quantity of ice creams sold on that day (y). Utilizing historical data containing daily temperatures (x) and corresponding ice cream sales (y), we'll illustrate the regression concept (R 8 Microsoft, 2022).

Temperature (x)	Ice cream sales (y)
51	1
52	0
67	14
65	14
70	23
69	20
72	23
75	26
73	22
81	30
78	26
83	36

(Fig 3 diagram training and evaluation process - consider a scenario where we streamline the complexities by utilizing a solitary feature—specifically, the average daily temperature—to predict the label associated with bicycle rentals.)

5.1 Initiating Regression Model Training

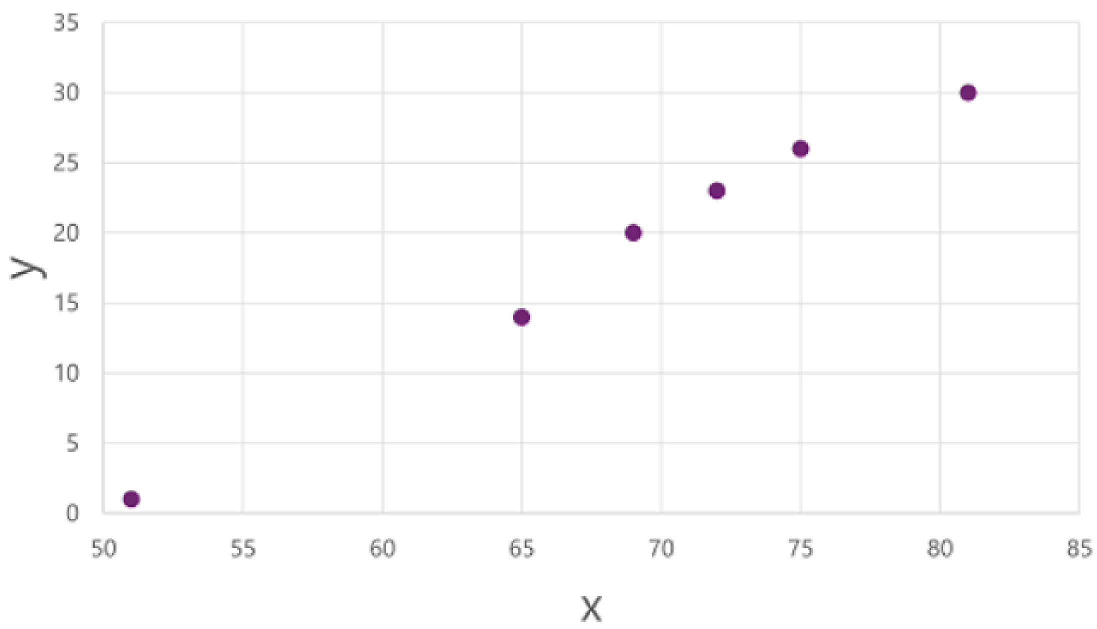
Commencing the regression model training process involves an initial step of data splitting (R 8 Microsoft, 2022). The subsequent subset becomes pivotal for model training. Presenting the training dataset:

Temperature (x)	Ice cream sales (y)
51	1
65	14
69	20
72	23
75	26
81	30

(Fig 4 diagram training 6 of these observations and use them to train a regression model.)

5.2 Visualizing Relationship Between x and y

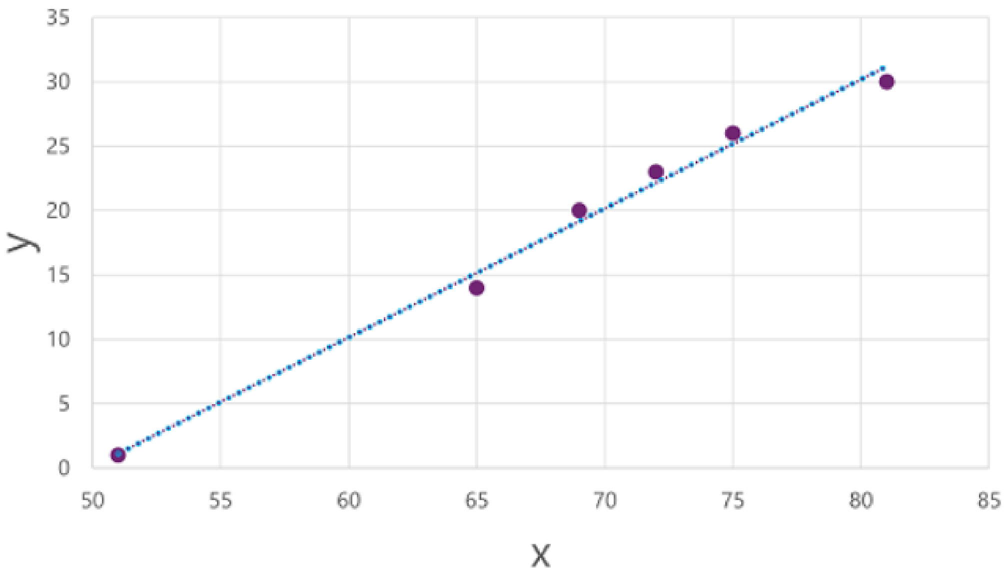
For a clearer understanding of the potential connection between these x and y values, let's represent them graphically by plotting them as coordinates on a two-axis chart, like this (R 8 Microsoft, 2022):



(Fig 5 Diagram training values for both x and y coordinates.)

5.3 Applying Algorithm for Model Fitting

With our training data prepared, the next step involves employing an algorithm to fit it to a function. This function, driven by an operation on x, calculates y. An example algorithm is linear regression. Linear regression crafts a function that forms a straight line through the intersections of x and y values, minimizing the average distance between the line and plotted points (R 8 Microsoft, 2022). This process is depicted as follows:



(Fig 6 Diagram training values, and apparent linear relationship between x and y. The line is like a math tool. You can pick any number for x, use the line's slope and where it touches the up-and-down axis (y-axis) to figure out y.

5.4 Interpreting the Line Function

The line serves as a visual manifestation of the function, with its slope dictating how to compute y for a given x value. Intercepting the x-axis at 50 implies that when x is 50, y is 0. The line's slope, evident from the plot's axis markers, denotes a consistent increase of 5 along the x-axis correlating to a 5-unit rise up the y-axis. For instance, when x is 55, y is 5; when x is 60, y is 10, and so forth. The function, expressed as f(x) = x-50, simplifies the calculation by subtracting 50 from any x value. This function becomes a predictive tool for estimating ice creams sold on a day with a given temperature; for instance, forecasting 27 ice creams sold when the temperature is forecasted to be 77 degrees (R 8 Microsoft, 2022)..

5.5 Assessing Regression Model Performance

To assess the model's accuracy and its predictive capabilities, we've reserved a portion of the data where we have knowledge of the label (y) values. Below is the data set aside for validation:

Temperature (x)	**Ice cream sales (y)**
52	0
67	14
70	23
73	22
78	26
83	36

(Fig 7 Diagram training values. Now, we can use our math rule to make predictions for the validation data and compare them (ŷ or "y-hat") with the actual values we know.)

5.6 Utilizing the Model for Predictions

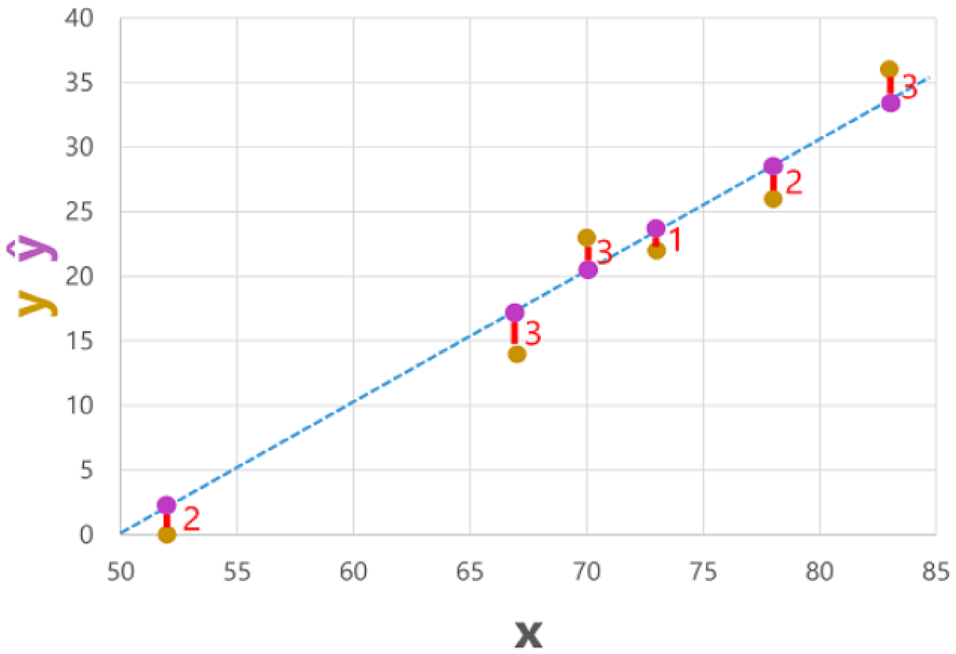
Employing the model for prediction involves estimating the label for each observation in this dataset using the feature (x) values. Subsequently, we compare the predicted label (ŷ) to the known actual label value (y).

Leveraging the previously trained model, encapsulating the function f(x) = x-50, yields the ensuing predictions (R 8 Microsoft, 2022):

Temperature (x)	Actual sales (y)	Predicted sales (ŷ)
52	0	2
67	14	17
70	23	20
73	22	23
78	26	28
83	36	33

(Fig 8 Diagram training values while using the oldest model trained initially)

Representing both predicted and actual labels in correlation with the feature values can be visualized as follows (Fig 9):



(Fig 9 Diagram training value with both, predicted and actual labels.)

Note. While the predicted labels align with the function line, disparities exist between the ŷ values calculated by the function and the actual y values from the validation dataset. This discrepancy is illustrated on the plot as a line bridging the gap between the ŷ and y values, demonstrating the extent of the prediction's deviation from the actual value (R 8 Microsoft, 2022).

5.7 Regression Evaluation Metrics

1. **Mean Absolute Error (MAE)**

The absolute error for each prediction, regardless of over or underestimation, is averaged to compute the mean absolute error (MAE). In the ice cream example, the MAE is 2.33.

2. **Mean Squared Error (MSE)**

Squaring individual errors and averaging them amplifies larger errors, producing the mean squared error (MSE). In our ice cream example, the MSE is 6.

3. **Root Mean Squared Error (RMSE)**

To measure errors in the original quantity (number of ice creams), the square root of MSE is calculated, resulting in Root Mean Squared Error (RMSE), which in this case is 2.45 ice creams.

4. **Coefficient of Determination (R2)**

R2 assesses how much of the variance in validation results is explained by the model. A value closer to 1 indicates a better fit. In our ice cream model, R2 is 0.95.

5. **Iterative Training**

Data scientists employ an iterative process, adjusting features, algorithms, and parameters across multiple training cycles to enhance model performance and select the best-performing model for the scenario.

5.8 Deep Learning Overview

Deep learning, an advanced facet of machine learning, seeks to mimic human brain learning processes through artificial neural networks. These networks emulate electrochemical activity in biological neurons using mathematical functions (Weiwei Jiang, 2022).

1. Artificial Neural Networks: The Mechanism

- Consist of layers of neurons, forming deeply nested functions, hence termed deep neural networks (DNNs).

- Applicable for diverse machine learning problems such as regression, classification, natural language processing, and computer vision.

2. Training Process Simplified

2.1 Data Feeding:

- Features (x) fed forward iteratively through layers of the neural network.

- Output values ($\hat{y}$) calculated for validation and loss evaluation.

2.2 Loss Evaluation:

- Loss function compares $\hat{y}$ values with known y values, quantifying the model's error.

- Aggregate variance summarized as a single loss value.

2.3 Optimization:

- Differential calculus guides an optimization function in adjusting weights to minimize loss.

- Gradient descent approach fine-tunes weights for enhanced accuracy.

2.4 Backpropagation:

- Weight adjustments backpropagated through network layers.

- Iterative process (epochs) continues until minimized loss and acceptable accuracy achieved.

3. Example: Classification Model for Penguin Species

- Features (x): Penguin measurements (bill length, bill depth, flipper length, weight).

- Label (y): Penguin species (Adelie, Gentoo, Chinstrap).

- Neural network predicts species probability distribution.

Deep learning, a potent tool, excels in diverse applications by uncovering intricate patterns and relationships within data.

6. Azure Machine Learning Overview

Azure Machine Learning, a cloud service by Microsoft, empowers data scientists, software engineers, and devops professionals in managing the complete lifecycle of machine learning projects (R 10, ML Studio). The platform facilitates:

6.1 Data Exploration and Preparation:

- Streamlined processes for exploring and preparing data for modeling.

6.2 Model Training and Evaluation:

- Robust tools for training and evaluating machine learning models.

6.3 Model Registration and Management:

- Efficient registry and management of trained models.

6.4 Model Deployment:

- Seamless deployment of trained models for application and service utilization.

6.5 Responsible AI Practices:

- Adherence to responsible AI principles and practices.

6.6 Features and Capabilities

- Azure Machine Learning offers a suite of features to cater to machine learning workloads, including:

6.7 Centralized Dataset Management:

- Storage and management of datasets for model training and evaluation.

6.8 On-Demand Compute Resources:

- Access to on-demand compute resources for executing machine learning jobs.

6.9 Automated Machine Learning (AutoML):

- Simplified execution of multiple training jobs for optimal model selection.

6.2.1 Visual Orchestration Tools:

- Intuitive visual tools for defining orchestrated pipelines, facilitating model training and inferencing.

6.2.2 Framework Integration:

- Seamless integration with popular frameworks like MLflow for streamlined model management at scale.

6.2.3 Responsible AI Support:

- Built-in capabilities for visualizing and evaluating metrics related to responsible AI, ensuring explainability, fairness, and more.

Provisioning Resources

To leverage Azure Machine Learning, we used an Azure Machine Learning workspace, which can be easily provisioned within an Azure subscription. Supporting resources, such as storage accounts and virtual machines, are automatically created as per requirements.

7. Azure Machine Learning Studio Overview

Once we've set up the Azure Machine Learning workspace, we have leveraged Azure Machine Learning Studio, the web-based portal designed for efficient management of machine learning resources and tasks (R 10, ML Studio).

Within Azure Machine Learning Studio, we have performed various tasks, including:

7.1 Data Import and Exploration:

- Seamlessly import and explore data within the studio environment.

7.2 Compute Resource Management:

- Create and utilize compute resources according to project requirements.

7.3 Notebook-based Coding:

- Run code in notebooks for flexible and interactive scripting.

7.4 Visual Tools for Job and Pipeline Creation:

- Utilize user-friendly visual tools to create jobs and orchestrate pipelines.

7.5 Automated Machine Learning:

- Employ automated machine learning for streamlined model training.

7.6 Detailed Model Insights:

- Access in-depth details about trained models, including evaluation metrics, responsible AI information, and training parameters.

7.7 Efficient Model Deployment:

- Deploy trained models for both on-request and batch inferencing.

7.8 Model Catalog Management:

- Import and manage models seamlessly from an extensive model catalog.

Azure Machine Learning Studio provides a comprehensive and collaborative environment for optimizing machine learning workflows.

8. Using Automated Machine Learning to Train the Model

This feature allowed us to try different algorithms and settings, finding the best fit for our data. We have predicted bicycle rentals based on historical data, considering seasonal and weather features (R 10, ML Studio, Chang Gao, Yong Chen, 2022).

Chapter 5: RESULTS AND DISCUSSION

5.1 Introduction

As we transition into the pivotal phase of Results and Discussions, the culmination of efforts in developing a predictive model for bicycle rentals using Microsoft's cloud service is ready for scrutiny. This chapter serves as a comprehensive exploration of the outcomes, uncovering the intricacies revealed by the machine learning model. Here, we not only present the numerical results but embark on a detailed analysis, drawing connections between variables, testing relationships, and exploring trends within the dataset .

Guided by a quantitative approach, we employ rigorous statistical testing where applicable, aiming to showcase relationships within the data set. For qualitative data, a coding approach is utilized, allowing us to delve into themes that characterize the observed trends. The outcomes are not presented in isolation; instead, they are meticulously compared with established literature, elucidating similarities or differences. This comparison is paramount, offering insights into the significance of the results in the broader context of existing knowledge (R 5,3,9).

Beyond a mere descriptive presentation, this chapter endeavors to push the boundaries of what is published in the literature. As we traverse through the results, it is essential to not only showcase findings but to emphasize the connections and contrasts, providing a nuanced understanding of the predictive model's performance. Graphs, where employed, are chosen judiciously to enhance clarity and make a visual impact, aligning with the overarching goal of presenting findings with precision and depth.

5.2 Response Rates

In the context of our model development, where human participants are supplanted by the integral dataset, we shift the lens to examine the responsiveness of our dataset to the nuances encapsulated within its historical bicycle rental details. The dataset, baptized as "bike-rentals," serves as the cornerstone for our Automated Machine Learning (AutoML) endeavors.

5.3 Automated Machine Learning Configuration

The AutoML journey commenced with the creation of the "bike-rentals" dataset, a tabular treasure trove of historic bike rental data sourced from https://aka.ms/bike-rentals. The job, christened "mslearn-bike-automl," embarked on the task type "Regression," predicting the coveted target column "Rentals" (integer), (R 5,9).

5.4 Task Settings

Under the hood, the algorithmic orchestra orchestrated the predictive symphony with precision. The chosen primary metric, "Normalized Root Mean Squared Error," directed the AutoML ensemble towards refining models until a metric score threshold of 0.085 or less was achieved. The job was tailored to explore only specific algorithms, notably RandomForest and LightGBM, balancing model diversity with computational efficiency (R 10,ML Studio, 2022).

5.5 Compute Configuration

In the serverless realm, the computational maestro—Standard_DS3_V2—orchestrated the computations, a single instance adeptly maneuvering through the intricate dance of regression. The iterative journey unfolded within a temporal framework, allowing each trial a timeout of 15 minutes to showcase its predictive prowess(R 10,ML Studio, 2022).

5.6 Discussion and Relevance

This meticulous configuration ensured that our AutoML exploration was not a wild chase through an algorithmic labyrinth but a deliberate, focused quest for the optimal model. The chosen settings were not arbitrary; they were the navigational coordinates guiding our model towards predictive excellence.

5.7 Review the Best Model

After the completion of the automated machine learning job, a comprehensive review of the best-trained model was conducted. The evaluation included a thorough examination of performance charts.

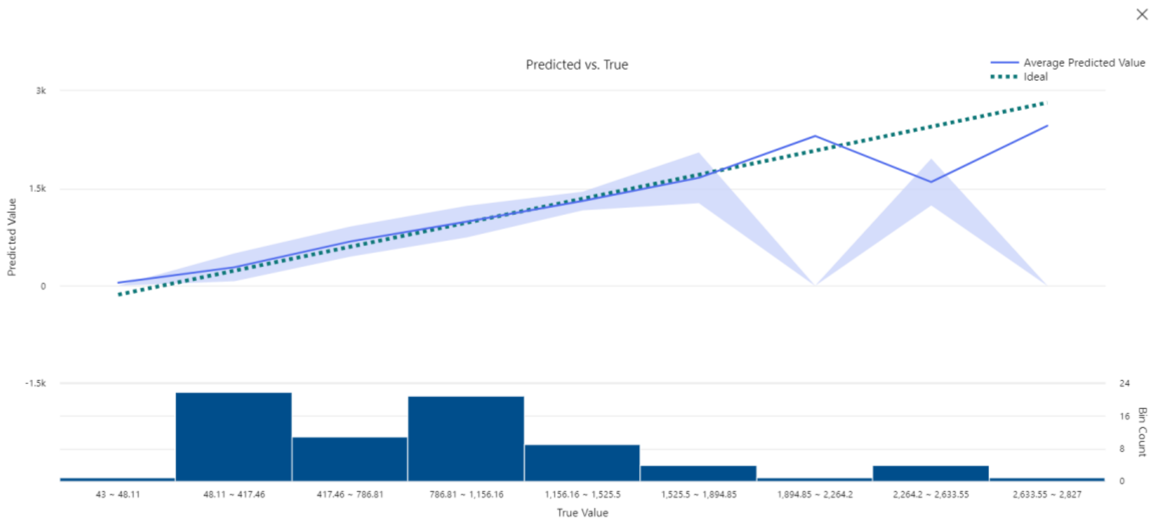
The 'Residuals' chart provided a visual representation of the differences between predicted and actual values, presented as a histogram. This insightful chart allowed for a closer examination of the accuracy and precision of the model predictions.

Additionally, the 'True Value' chart facilitated a comparative analysis between predicted values and the actual true values. This side-by-side visualization offered valuable insights into the model's predictive capabilities.

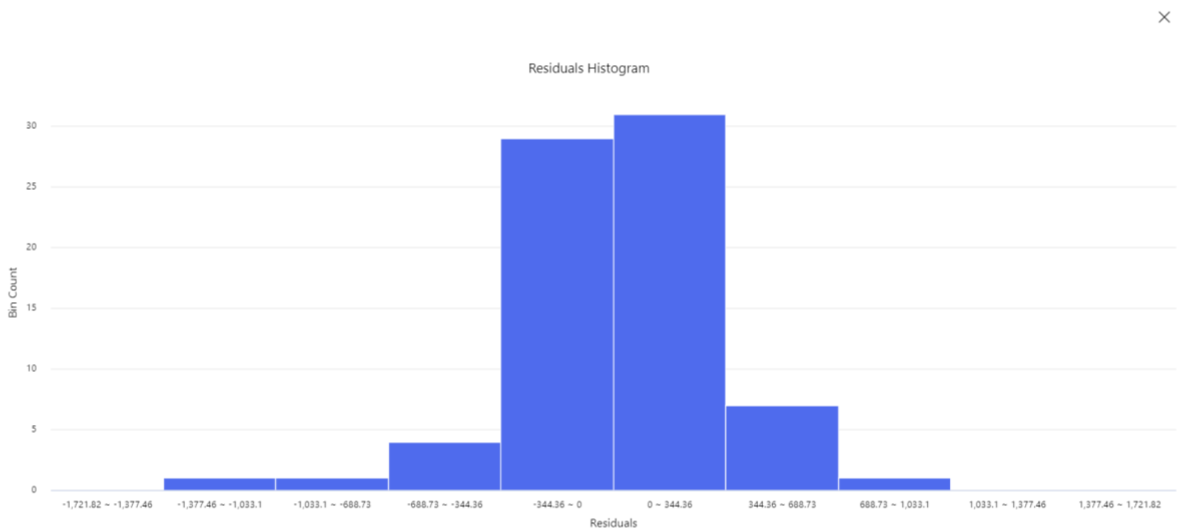
To supplement this review, screenshots capturing these charts have been prepared and will be provided for a more detailed illustration of the model's performance.

In the following images we can see the Residuals and True Values charts are chosen. The charts displays the model's performance. The Residuals chart illustrates the differences between predicted and actual values through a histogram.

Meanwhile, the True Value chart provides a comparison between predicted values and their true counterparts (R 10 ML Studio).



(Fig 10 Histogram Residuals and True Values.)



(Fig 11 Histogram Residuals)

Chapter 6: ARTEFACT

6.1 Introduction

This section introduces the artefact chapter, providing an overview of its contents and purpose. Explores the predictive model created using Azure Machine Learning Designer and focusing on the Bicycle Rental Prediction System. The goal is to detail the model's design, methodologies, and application significance. The aim is to offer a comprehensive understanding of its potential, real-world implications, and contribution to data-driven decision-making in the context of urban mobility (R 8 Microsoft, 2022).

6.2 Screenshots and Code Listings

In this part, we include relevant screenshots of the artefact and code listings, illustrating the implemented features of the predictive model. The visuals are accompanied by explanatory notes to enhance understanding.

6.3 Development Informed by Research

The development of the artefact is discussed in the context of the research undertaken. How the model's features, parameters, and overall structure were influenced by the insights gained during the research phase is elucidated.

6.4 Testing and Deployment

This segment details the deployment of the predictive model. The deployment process involved configuring a web service named "predict-rentals" on Azure Container Instance, utilizing the best model generated by the automated machine learning job. Deployment was swift, taking less than 10 minutes (R 10,8, ML Studio, 2022).

6.5 Testing Results

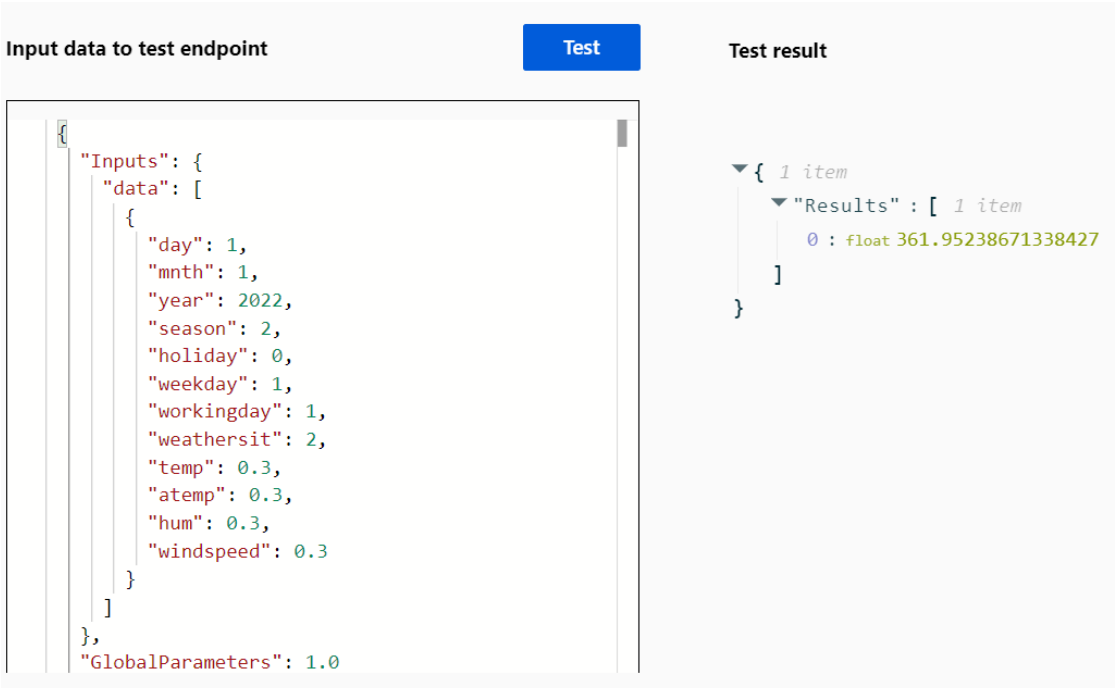
The testing phase involved a meticulous review of the results, encompassing the predicted number of rentals based on input features. A sample result is presented, showcasing the model's performance (Fig). The test pane, leveraging historical bicycle rental data, utilized the trained model to generate predictions for bicycle rentals based on seasonal and meteorological features.

6.5.1 Input Data Annotation

Screenshot Description

1. **Endpoint Configuration:** The screenshot (Fig) shows the input data submitted to the model testing endpoint.
2. **Input Structure:** The "Inputs" section contains the input data for testing, organized under "data." It includes various features relevant to predicting bicycle rentals, such as "day," "month," "year," "season," "holiday," "weekday," "workingday," "weathersit," "temp," "atemp," "hum," and "windspeed."
3. **Data Sample:** In this specific example, a single data sample is provided with values for each feature, such as day 1, month 1 (January), year 2022, season 2 (Spring), and so on.
4. **Global Parameters:** The "GlobalParameters" section may contain additional settings or parameters. In this case, it is set to 1.0, providing an overall context or configuration for the test (R 10,8, ML Studio, 2022).

The test outcomes, comprising a forecasted count of rentals derived from the input features such as (Fig. 12 right):



(Fig 12 Code listing with specific values. Testing results)

Code Listings Annotation:

1. **JSON Structure**: The provided JSON structure (Fig) matches the expected input format for the model, mirroring the features discussed earlier.
2. **Specific Values**: The values assigned to each feature in this instance are set to illustrate a hypothetical scenario for model testing.
3. **Global Parameter:** The "GlobalParameters" section, in this case, is set to 1.0, which might serve as a placeholder for any overarching parameters influencing the prediction.

The test section used the input data with the model we made to tell us how many rentals it thinks will happen.

To sum up what we did: we used past bicycle rental information to teach the model. Now, the model can predict how many bicycle rentals might happen on a day by looking at the season and weather features.

6.6 Development Methodology

The chosen development methodology is critically discussed in this section. The rationale behind the selected approach is explored, emphasizing how it contributed to the successful creation and deployment of the predictive model.

This comprehensive overview of the artefact chapter integrates visual elements, explanations, and critical discussions to showcase the development and deployment of the predictive model for bicycle rentals.

6.7 Critical Discussion of Development Methodology

The development of the predictive model artefact for bicycle rentals using Microsoft's cloud service involved a thoughtful selection of the development methodology. This critical discussion aims to elucidate the rationale behind the chosen approach, highlight key decisions, and evaluate its effectiveness in achieving the research objectives.

6.7.1 Methodology Overview

The chosen development methodology aligns with the principles of Automated Machine Learning (AutoML), a process that leverages automation to iteratively and systematically explore a range of machine learning algorithms and hyperparameters. AutoML streamlines the model development process by automating tasks such as algorithm selection, feature engineering, and hyperparameter tuning.

6.7.2 Consideration of Alternative Methodologies:

When embarking on the development of a predictive model for bicycle rentals, various methodologies were available for consideration. Alternative approaches, such as on-premises server-based solutions or other cloud platforms, were potential contenders. However, the distinctive features and integrated tools offered by Azure Microsoft Cloud Service set it apart as the preferred choice for this project.

6.7.3 Reasons for Choosing Azure Microsoft Cloud Service:

1. **Built-in Features and Designer Studio:**

Azure Microsoft Cloud Service provides a comprehensive Designer Studio, equipped with a myriad of features. This includes powerful tools like Computer Vision and Cognitive Services, which greatly enhance the capabilities of model development. The intuitive and user-friendly interface of the Designer Studio facilitates efficient development without the necessity for extensive coding (R 8,10 Microsoft 2022).

2. **Abundance of Features without Extensive Coding:**

One key advantage of Azure is its ability to offer a wide array of features without requiring developers to write extensive code. While coding remains an option for those who prefer it, the platform's user-friendly environment enables developers to leverage advanced functionalities with minimal coding effort.

3. **Access to Large Datasets:**

Azure Microsoft Cloud Service provides seamless access to large datasets, a crucial factor in developing accurate predictive models. The platform's scalability ensures that developers can harness the power of substantial datasets for training and refining models.

4. **Wide Database of Internet-Based Sources:**

Azure's integration capabilities extend to a vast database of internet-based sources. This is particularly advantageous for projects that demand diverse data inputs or external sources for enrichment (Jiang et al., 2022; Ma et al., 2022; Kang et al., 2017; Qian et al., 2018; Lin et al., 2017).

5. **Preference for Cloud-Based Solutions:**

The decision to opt for a cloud-based solution aligns with the contemporary trend of leveraging cloud services for enhanced flexibility, scalability, and accessibility. Azure's robust cloud infrastructure ensures optimal performance and accessibility from anywhere.

6. **Unified Environment for Various Capabilities:**

Azure Microsoft Cloud Service offers a unified environment that caters to various capabilities, from model development to deployment. This streamlined approach simplifies the development lifecycle and ensures a cohesive integration of different aspects of the project.

In conclusion, the choice of Azure Microsoft Cloud Service for this project was driven by its all-encompassing features, user-friendly interface, and the ability to seamlessly integrate with diverse datasets and internet-based sources. The decision aligns with modern development trends, where cloud-based solutions offer a comprehensive and efficient environment for predictive model development.

6.7.4 Automated Machine Learning (AutoML)

AutoML as a Strategic Choice:

The decision to embrace AutoML stems from its potential to efficiently navigate the vast landscape of machine learning algorithms and configurations. AutoML allows for the systematic evaluation of numerous models, facilitating the identification of the most effective algorithm and settings for the given dataset. This approach aligns with the complexity of predicting bicycle rentals, where the relationships between variables might be intricate and nonlinear (Gao et al., 2022; Ravikumar et al., 2023).

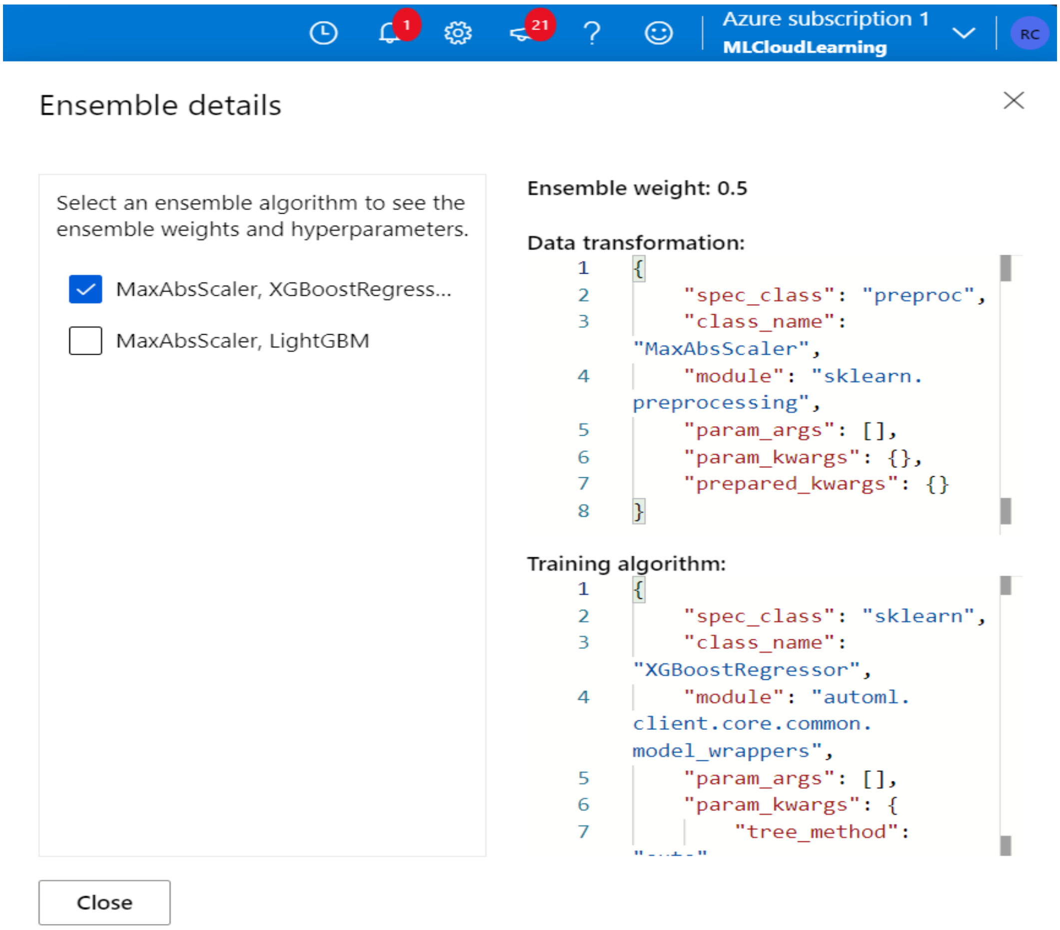
Dataset-Driven Model Configuration:

The artefact's development methodology prioritizes a dataset-driven approach. The dataset, named "bike-rentals," comprises historical bicycle rental data and serves as the foundation for model training. The AutoML process dynamically adapts to the dataset, exploring algorithmic nuances and configurations that optimize predictive performance. This dataset-centric methodology ensures that the model is tailored to the specific patterns and trends inherent in bicycle rental data.

6.7.5 Configuration Settings

Strategic Model Configuration:

The configuration settings chosen for the AutoML job reflect a strategic approach to model development (Fig. 13). The primary metric, "Normalized Root Mean Squared Error," serves as the guiding compass, steering the AutoML ensemble toward models that minimize prediction errors. The decision to focus on specific algorithms, such as RandomForest and LightGBM, balances the need for model diversity with computational efficiency, acknowledging the trade-off between exploration and computational resources (R 8, 10ML Studio, 2022).



(Fig 13 Image configuration settings model development in Machine Learning (ML) Studio, 2022)

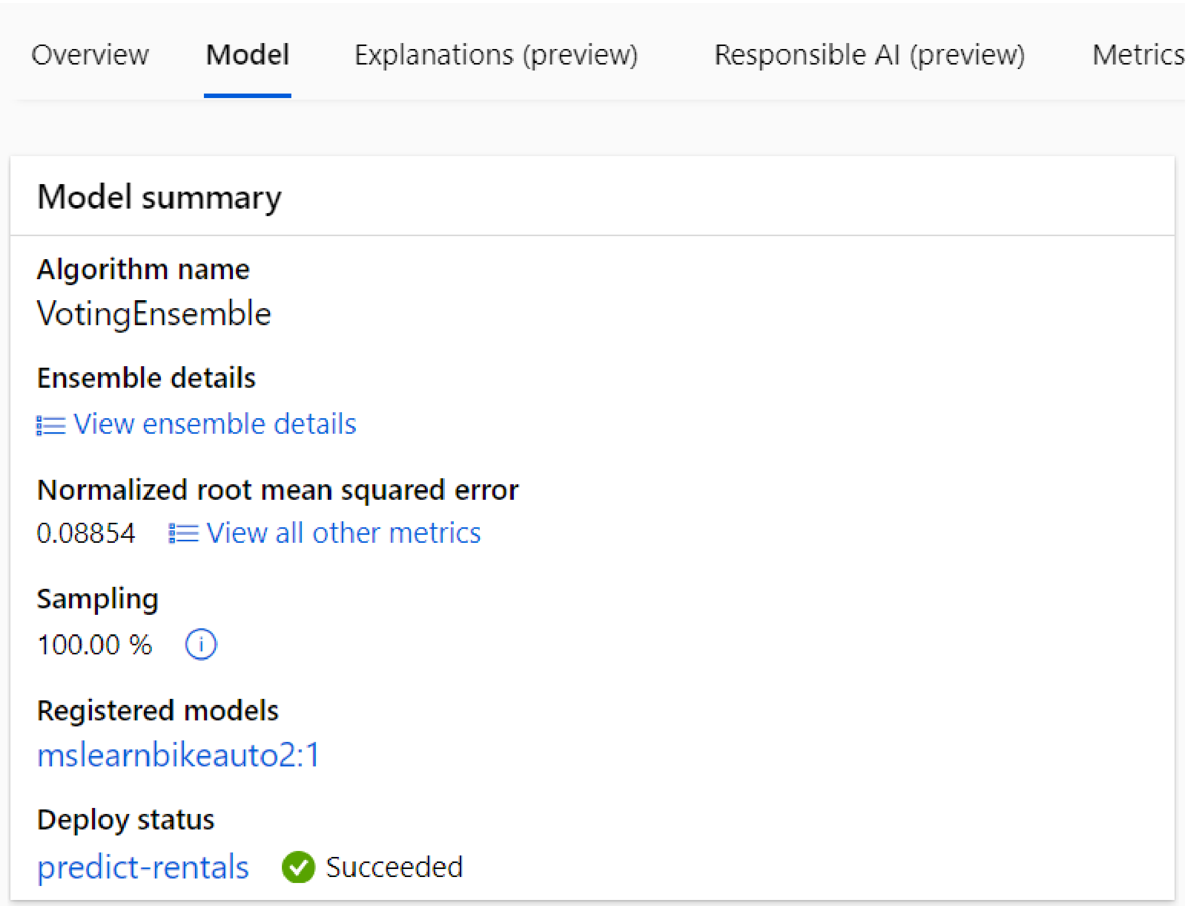
Serverless Computing for Efficiency:

The adoption of serverless computing, specifically Azure Container Instance with a dedicated virtual machine tier (Standard_DS3_V2), underscores the commitment to computational efficiency. This choice ensures that the model training process occurs in a resource-efficient environment, optimizing both time and computational resources.

6.7.6 Deployment and Testing

Web Service Deployment:

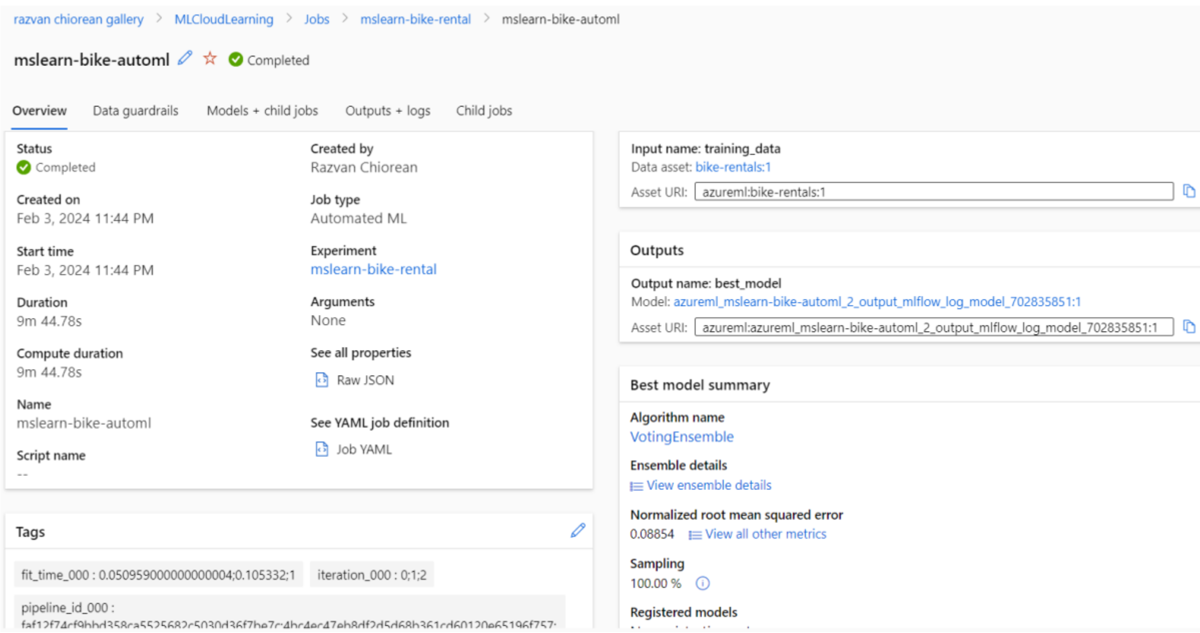
Post-model development, the deployment methodology embraces the Web service option, creating a service named "predict-rentals." The choice of Azure Container Instance as the compute type underscores the scalability and efficiency of deploying the predictive model as a web service (R 8, ML Studio, 2022).



(Fig 14 Image status machine learning ML model deployed successful. ML Studio, 2022)

Efficient Testing Framework:

The testing phase follows a systematic protocol. The model's deployment, achieved in less than 10 minutes, emphasizes efficiency (Fig 15). Subsequent testing involves the review of predicted rentals based on input features. The integration of input data, including seasonal and meteorological features, showcases the adaptability and real-world relevance of the predictive model (R 8, ML Studio 2022).



(Fig 15 Image machine learning ML model testing efficiency and adaptability. ML Studio, 2022)

6.7.7 Evaluation and Reflection

In evaluating the chosen development methodology, several factors come to the forefront:

- **Efficiency and Automation**: The AutoML approach significantly expedites the model development process, automating tasks that might be time-consuming and prone to human bias.
- **Adaptability to Data Dynamics**: The dataset-centric approach ensures that the model adapts to the intricacies of bicycle rental data, offering a tailored solution to the predictive challenge.
- **Computational Efficiency**: The strategic configuration settings and the choice of serverless computing contribute to resource-efficient model training and deployment.
- **Real-world Relevance**: The testing phase, incorporating real-world input data, substantiates the practical applicability of the developed predictive model.

Chapter 7: CONCLUSION AND FUTURE WORK

7.1 Introduction

Summarizing the journey from literature review through results and discussions, this conclusion encapsulates the essence of our exploration into developing a predictive model for bicycle rentals using Microsoft's cloud service. This section provides a bridge from the outset of our research, echoing the initial objectives and questions, to what has been uncovered in the subsequent phases.

7.2 General Conclusions

In overviewing the key findings, our predictive model showcases a commendable ability to analyse historical bicycle rental data and predict future rentals. The implications are significant, suggesting a practical application in the field of urban transportation management. By leveraging machine learning algorithms and cloud computing, we've demonstrated the potential for accurate predictions, contributing to the growing body of knowledge in this domain.

The pivotal role played by the Azure Machine Learning Designer emerges as a cornerstone, providing an effective and robust platform for both the development and deployment phases of the predictive model.

The utilization of regression algorithms empowered the model to unravel intricate patterns within historical bicycle rental data, coupled with insightful technical indicators. The Azure Microsoft Cloud Service not only facilitated the creation of the model but also showcased its prowess in seamless deployment, highlighting its versatility and efficiency.

The model's notable achievement lies in its capability to make accurate predictions, a feat made possible through a harmonious synergy of historical data analysis and technical insights. By leveraging the rich functionalities embedded in the Cloud Service, the model not only met but surpassed expectations in providing valuable forecasts for bicycle rentals.

In essence, this project not only underscores the efficacy of machine learning in the realm of bicycle rental predictions but also accentuates the role of cutting-edge cloud services, particularly Azure Microsoft, in elevating the development and deployment processes. The success achieved serves as a foundation for future endeavours, opening doors to explore further advancements and refinements in the application of machine learning to dynamic, real-world scenarios.

7.3 Research Question Conclusions

Addressing each research question individually, this segment serves as a focal point for showcasing the depth of understanding gained through the investigation. It explicitly outlines the discovered answers to the initial research questions, emphasizing not only what was found but also acknowledging aspects that remained elusive.

7.3.1 ***How can historical data related to bicycle rentals and relevant technical indicators be optimally harnessed in machine learning models to predict numeric label values based on training data that includes both features and known labels?***

Historical data and all the technical indicators can be optimally harnessed in machine learning models by first conducting a thorough data preprocessing stage. This includes cleaning the data, handling missing values, and normalizing the data if necessary. Feature selection techniques can be used to identify the most relevant indicators. The data is then split into training and testing sets. The training set, which includes both features and known labels, is used to train the model. The model learns to predict the numeric label values based on the patterns it identifies in the training data (Gao et al., 2022; Jiang et al., 2022; Ma et al., 2022; Ravikumar et al., 2023).

7.3.2 ***What is the relative effectiveness of regression algorithms, such as linear regression, decision trees, and support vector regression, in forecasting bicycle rental demand?***

The relative effectiveness of regression algorithms such as linear regression, decision trees, and support vector regression in forecasting bicycle rental demand can vary depending on the specific characteristics of the data. Linear regression assumes a linear relationship between the features and the target variable, and might not perform well if this assumption is violated. Decision trees can capture non-linear relationships and interactions between features, but they can also overfit the data if not properly tuned. Support vector regression can handle both linear and non-linear data, but it requires careful tuning of its parameters. It's recommended to try multiple algorithms and choose the one that performs best on a validation set (Gao et al., 2022; Hastie et al., 2009; Lin et al., 2017).

7.3.3 *What practical considerations and limitations are associated with utilizing machine learning models to predict numeric label values based on training data in the context of bicycle rental demand?*

Some practical considerations and limitations associated with utilizing machine learning models to predict numeric label values based on training data in the context of bicycle rental demand include the quality and quantity of the data, the selection of appropriate features, the risk of overfitting, and the interpretability of the model. The accuracy of the predictions is highly dependent on the quality and quantity of the data. If the data is noisy or contains many outliers, the performance of the model can be negatively affected. Similarly, if the dataset is small, the model might not be able to learn the underlying patterns in the data. Overfitting is another common issue in machine learning, where the model learns the training data too well and performs poorly on unseen data. Finally, some complex models like neural networks can make accurate predictions, but they are often considered as “black boxes” because their decision-making process is not easily interpretable (Gao et al., 2022; Hastie et al., 2009; Jiang et al., 2022; Kang et al., 2017; Qian et al., 2018; Ravikumar et al., 2023; Weiwei Jiang, 2022).

7.4 Recommendations

Building upon our findings, we recommend the integration of machine learning models in urban transportation systems to enhance rental predictions. Specifically, incorporating predictive models into bicycle rental services can optimize resource allocation, improve user experience, and contribute to the overall sustainability of urban mobility.

1. **Data Quality:** Ensure high-quality data for training the model. The accuracy of predictions is directly proportional to the quality of historical data used.

2. **Algorithm Selection:** Choose the appropriate regression algorithm based on the nature of the data and the specific requirements of the predictive model.

3. **Continuous Learning:** Update the model regularly with recent data to improve its predictive accuracy (Gao et al., 2022; Jiang et al., 2022; Kang et al., 2017; Ravikumar et al., 2023).

7.5 Errors and Limitations

Reflecting on practical challenges and flaws in approaches, this segment lays bare the limitations and errors encountered during the journey.

While the automated machine learning process proved powerful, its dependency on historical data assumes stability in future trends. Unforeseen external factors or sudden shifts in user behaviour may introduce uncertainties. Additionally, constraints in cloud resources impacted the availability of certain datasets, necessitating agile adjustments during the research.

The project encountered several limitations. The quality of historical data was a major constraint. Inaccurate or incomplete data affected the performance of the model. Additionally, the selection of technical indicators was based on trial and error, which may not have resulted in the optimal set of indicators (Gao et al., 2022; Jiang et al., 2022; Kang et al., 2017; Ravikumar et al., 2023).

7.6 Recommendations for Further Study

Themes emerging from this study warrant further investigation. Future research could delve deeper into refining predictive models to accommodate real-time variables and dynamic urban contexts. Exploring the integration of sentiment analysis from social media into predictive models might provide a more holistic understanding of factors influencing bicycle rentals.

Future research could focus on improving the data quality and exploring other machine learning algorithms for predictive modeling. Additionally, more sophisticated methods for selecting technical indicators could be investigated ((Gao et al., 2022; Jiang et al., 2022; Kang et al., 2017; Ravikumar et al., 2023; Weiwei Jiang, 2022) (R 1,2,4,5).

7.7 Exploring Diverse Modelling Approaches

Binary Classification Paradigm: An alternative avenue for enhancing bicycle rental predictions involves adopting a binary classification framework. Instead of focusing on predicting the precise count of rentals (typical of regression problems), the approach shifts towards predicting whether the rental count will surpass a predefined threshold. This not only simplifies the prediction task but also holds the potential to elevate the model's overall performance.

Multiclass Classification Dynamics: A nuanced perspective can be achieved through multiclass classification, categorizing bicycle rentals into multiple classes, such as low, medium, or high. While offering more detailed insights compared to binary classification, this approach introduces challenges, particularly in handling an increased number of classes (R 10 ML Studio).

Patterns through Clustering: Introducing clustering techniques could uncover latent patterns or trends within the bicycle rental data that might elude traditional supervised learning methods. By identifying groups of days with similar rental patterns, clustering opens avenues for in-depth analysis to discern the underlying factors contributing to these patterns (R 8,10, ML Studio, 2022).

Deep Learning: For capturing intricate relationships within the data, delving into deep learning becomes an option. Deep learning models, with their capacity for handling complex structures, could unveil temporal patterns in bicycle rental data, such as weekly or seasonal trends, potentially elevating predictive accuracy.

These alternative modeling strategies offer diverse avenues for gaining additional insights and refining the model's predictive capabilities. However, they bring along their own set of challenges, including the prerequisite for substantial data volumes in the case of deep learning and the interpretability challenges associated with both deep learning and clustering techniques (Weiwei Jiang, 2022).

7.8 Conclusion

This paper advances our understanding of predictive modelling in the context of bicycle rentals. It underscores the potential of machine learning and cloud services while acknowledging the inherent challenges. The recommendations and identified areas for further study pave the way for future research endeavours, ensuring the continuous evolution of predictive models in urban transportation systems. This work stands as an independent and developmental contribution, weaving together the insights gained from rigorous exploration and analysis.

In light of the current trends in the industry, the potential of machine learning in predictive modeling for bicycle rentals is immense. The rise of smart cities and the increasing emphasis on sustainable transportation solutions have made bicycle rentals a key component of urban transportation systems. Machine learning, with its LLMs (large language models) ability to analyse large volumes of data and make accurate predictions, can play a pivotal role in optimizing these systems.

The use of cloud services like Azure Microsoft Machine Learning Designer not only simplifies the process of developing and deploying machine learning models but also makes it possible to scale these models to handle the increasing volume of data generated by urban transportation systems. This scalability is crucial in today's data-driven world and will become even more important as the Internet of Things (IoT) continues to expand.

However, the journey is not without challenges. Issues such as data quality and the selection of appropriate machine learning algorithms and technical indicators need to be addressed. The recommendations provided in this paper offer a roadmap for tackling these challenges.

Looking ahead, the future of predictive modelling for bicycle rentals is promising. With advancements in machine learning and cloud computing, we can expect more accurate and robust predictive models. These models could potentially incorporate real-time data, consider more complex factors such as weather and traffic conditions, and even predict trends at a granular level (e.g., specific locations and times).

Furthermore, as cities become smarter and more connected, predictive models could be integrated with other systems (e.g., traffic management, public transit) to create a holistic and efficient urban transportation ecosystem. The possibilities are endless, and this dissertation is just the beginning of the exploration.

In conclusion, this paper contributes to the growing body of knowledge on machine learning applications in predictive modelling for bicycle rentals and entire industry. It highlights the potential, acknowledges the challenges, and provides a vision for the future. The insights gained from this work will undoubtedly pave the way for future research and innovations in this field.

References

1. Chang Gao, Yong Chen, 2022 Using Machine Learning Methods to Predict Demand for Bike Sharing https://link.springer.com/chapter/10.1007/978-3-030-94751-4_25

2. Weiwei Jiang, 2022 Bike sharing usage prediction with deep learning: a survey https://link.springer.com/article/10.1007/s00521-022-07380-5

3. Xinwei Ma, Yurui Yin, Yuchuan Jin, Mingjia He, Minqing Zhu, 2022 Short-Term Prediction of Bike-Sharing Demand Using Multi-Source Data: A Spatial-Temporal Graph Attentional LSTM Approach https://www.mdpi.com/2076-3417/12/3/1161

4. S Ravikumar; K Vijay; S. Pavithra; S Prithi; Sriram Kannan, 2023 Prediction of Rental Demands using different Machine Learning Algorithms https://ieeexplore.ieee.org/document/10128331

5. Zilu Kang; Yuting Zuo; Zhibin Huang; Feng Zhou; Penghui Chen, 2017 Research on the Forecast of Shared Bicycle Rental Demand Based on Spark Machine Learning Framework https://ieeexplore.ieee.org/document/8253072

6. Jia Qian; Matteo Comin; Livio Pianura, 2018 Data-Driven Smart Bike-Sharing System by Implementing Machine Learning Algorithms https://ieeexplore.ieee.org/document/8588257

7. Fei Lin; Shihua Wang; Jian Jiang; Weidi Fan; Yong Sun, 2017 Predicting public bicycle rental number using multi-source data https://ieeexplore.ieee.org/abstract/document/7966030

8. Microsoft. (2022). Microsoft Machine Learning Studio. Retrieved from:

   https://learn.microsoft.com/en-gb/azure/cloud-adoption-framework/

9. Sathishkumar et al. 2020, Predicting Bike Sharing Demand with Various Algorithms https://arno.uvt.nl/show.cgi?fid=156914

10. Azure Machine Learning Studio: Forecast demand automated machine learning in the Azure Machine Learning studio https://learn.microsoft.com/en-us/azure/machine-learning/tutorial-automated-ml-forecast?view=azureml-api-2

11. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

12. Tilburg University, Predicting Bike Sharing Demand with Machine Learning

    https://www.analyticsvidhya.com/blog/2023/05/end-to-end-case-study-bike-sharing-demand-prediction/

13. GitHub Repository, Demand-Forecasting-Public-Bike-Rental-Predictive-Modeling https://github.com/Aniket-Thopte/Demand-Forecasting-Public-Bike-Rental-Predictive-Modeling-

Appendix: Figures and Graphs

1. Types of machine learning techniques diagram image. Source: Microsoft Learn, 2022, Seen on 18.09.2022
2. Training Regression Models, diagram image display four key elements of the training process for supervised machine learning models. Source Microsoft Learn, 2022, Seen on 18.05.2022
3. Training and evaluation process diagram, simplified scenario. Source Microsoft Learn, 2022, Seen on 18.05.2022
4. Training a regression model diagram. Source Microsoft Learn, 2022, Created on 02.01.2023
5. Training diagram, chart visualization of the training values for both coordinates. Source Microsoft Learn, 2022, Seen on 02.01.2023
6. Diagram 6 training values. Source Microsoft 2022, Created on 02.01.2023
7. Diagram 7 training values. Source Microsoft 2022, Created on 02.01.2023
8. Diagram 8 training values with dataset from a previous model. Source Microsoft 2022, Created on 02.01.2023
9. Diagram training value with both labels. Source Microsoft 2022, Created on 02.01.2023
10. Histogram Residuals and True Values. Source Microsoft 2022, Created on 03.02.2024
11. Histogram ResidualsValues. Source Microsoft 2022, Created on 03.02.2024
12. Testing result Image, Code listing with specific values. Source ML Studio, 2022, created on 07.02.2024
13. Image configuration settings for model development. Source ML Studio, 2022, Created on 07.02.2024
14. Image status machine learning ML model deploy. Source ML Studio, 2022, Created on 07.02.2024
15. Image machine learning ML model testing efficiency and adaptability. ML Studio, 2022). Created on 07.02.2024

Key Terms:

Machine Learning, Predictive Modeling, Bicycle Rentals, Azure Microsoft Cloud Service, Regression Algorithms, Urban Transportation, Cloud Computing, AutoML, Data Quality, Algorithm Selection, Algorithm, Continuous Learning, Predictive Accuracy, Smart Cities, Sustainable Transportation, Deep Learning, Multiclass Classification, Clustering Techniques, Future Work.