Predicting AI Service Focus in Companies Using Machine Learning: A Data Mining Approach with Random Forest and Support Vector Machine

Thosporn Sangsawang[1,*], Lin Tang[2], Tiamyod Pasawano[3]

[1,3]*Educational Technology and Communications Division, Faculty of Technical Education, Rajamangala University of Technology Thanyaburi, Thailand*

[2]*Vocational Education Division, Faculty of Technical Education, Rajamangala University of Technology Thanyaburi, Thailand*

(Received: January 5, 2024; Revised: March 10, 2024; Accepted: April 15, 2024; Available online: June 5, 2024)

Abstract

This study investigates the prediction of AI service focus in companies using machine learning models. The primary objective is to predict the percentage of AI service focus based on company characteristics such as project size, hourly rate, number of employees, and geographical location. Two machine learning models, Random Forest Regressor and Support Vector Regressor (SVR), were trained and evaluated to determine their effectiveness in predicting AI adoption. The dataset consists of 3099 companies, with key features cleaned and preprocessed, including the transformation of categorical variables into numerical ones using one-hot encoding and imputation techniques applied to handle missing values. The Random Forest model demonstrated better performance, with an R^2 value of 0.12, indicating a modest ability to explain the variance in AI service focus. In contrast, the SVR model had a negative R^2 value of -0.03, suggesting that it struggled to capture the underlying relationships in the data. The analysis identified project size and hourly rate as the most significant predictors of AI service focus, with larger projects and higher hourly rates correlating with a greater emphasis on AI services. Despite the relatively low performance of both models, this research provides valuable insights into the factors that influence AI adoption. The findings emphasize the importance of project-related characteristics in determining a company's AI service focus. However, the study is limited by missing data and the absence of additional features that could further improve prediction accuracy. Future research could benefit from incorporating more business-specific features and advanced modeling techniques to enhance the predictive power and generalizability of the model.

Keywords: AI Service Focus, Machine Learning, Random Forest, Support Vector Regressor, Feature Importance

1. Introduction

The rising significance of Artificial Intelligence (AI) services in contemporary business practices is reshaping company operations across various sectors. As organizations recognize AI's capability to enhance efficiency and improve decision-making, they are increasingly integrating these technologies into their service frameworks. This shift reflects a fundamental transformation in how businesses operate, compete, and evolve in a digitized marketplace. A critical aspect of this evolution is the necessity for organizations to anticipate their focus on AI services, steering their strategies toward successful implementation and optimal utilization.

The integration of AI technologies dramatically alters the landscape of customer service and operational efficiency. According to Khan et al., the disruption caused by AI has become integral to firms' digital strategies as they strive for competitive advantage and relevance in increasingly saturated markets [1]. The ability of AI to analyze vast amounts of data and discern patterns significantly contributes to fostering innovation, optimizing operational capabilities, and enhancing customer engagement. Furthermore, Malsia and Loku emphasize the importance of understanding consumer perceptions and trust regarding AI, indicating that acceptance of AI in service delivery is crucial for its successful adoption within sectors like healthcare [2].

Moreover, companies are grappling with how AI will redefine employment and task structures. Huang and Rust articulate that while AI has the potential to replace transactional job roles, those requiring emotional intelligence and human interaction are likely to retain their human-centric jobs in the near future [3]. This dual effect of AI presents a

***Corresponding author: Thosporn Sangsawang (sthosporn@rmutt.ac.th)**
DOI: https://doi.org/10.47738/ijaim.v4i2.83

conundrum where leveraging AI can lead to enhanced efficiencies while simultaneously necessitating a workforce adept at working alongside machines. It suggests a shift towards nurturing intuitive and empathetic skills in contrast to being purely analytical, thereby redefining the skill set required in the job market [4].

This landscape is further complicated by the inherent challenges associated with AI adoption. Fiske et al. highlight the ethical implications surrounding AI use, especially in areas like mental health, where AI applications must be implemented carefully to avoid undermining traditional care systems [5]. There exists a balancing act for companies to ensure that AI enhances rather than diminishes the level of service provided, particularly in sectors that require a high degree of personal interaction. This indicates that organizational leaders must approach AI integration with a holistic strategy that encompasses customer needs and ethical considerations.

The four essential axes identified by Vilaginés for effective AI integration into business strategy—focused AI strategy, customer understanding, effective customer interaction, and efficient implementation of AI—produce a roadmap that organizations can follow to streamline their AI efforts [6]. By maintaining a deep understanding of customer behavior and needs, businesses can tailor their AI services to provide personalized experiences, thus increasing customer satisfaction and loyalty. Properly executed, this strategy can transform AI from a mere technological enhancement into a central pillar of organizational processes, enabling firms to not only meet but exceed customer expectations.

To successfully navigate the focus of AI services within their offerings, companies must engage in comprehensive market research and strategic foresight. Studies have indicated that organizations adopting AI are achieving operational efficiency while redefining their leadership structures to accommodate data-driven decision-making processes [7]. This shift towards a more collaborative and transparent style of leadership is essential as it empowers teams to harness AI's potential effectively.

In sectors like banking, Niroula and Adhikari report that AI applications have proven transformative, enhancing customer relationships and guiding complex financial decisions through smart automation and personalized services [8]. However, they underscore that organizations must remain vigilant regarding security challenges and privacy issues involved in these initiatives. Clearly, organizations must be prepared to implement robust security measures alongside their AI technologies to protect user data and maintain trust.

Ultimately, AI services present an immense opportunity for businesses to revolutionize their operational frameworks. As Tula et al. assert, businesses that strategically implement AI are not only enhancing customer experiences but also redefining the narrative of service provision itself [9]. This narrative is underpinned by ethical framework considerations, where a delicate balance must be struck between maximizing technological advantages and preserving human-centric service quality. The successful orchestration of AI services within business models indicates a forward-looking perspective that comprises ethical, operational, and strategic planning layers, critical for sustainable growth in this era of digital transformation.

In this complex landscape of AI integration, companies must emphasize the continuous evolution of their workforce through training and adaptation, as highlighted by Abhulimen and Ejike in their findings regarding the need for improving customer service and operational innovations through AI [10]. The investment in human resources to harmonize with AI technologies is crucial, fostering an agile environment that encourages both human and machine collaboration. In conclusion, as the importance of AI services continues to expand, organizations that effectively predict and adapt to these changes will not merely survive but thrive in the increasingly competitive and digitized global market.

Understanding the factors that determine the level of AI focus in businesses is crucial for organizations seeking to enhance their competitive advantage and adapt to the evolving market landscape. Companies are increasingly leveraging AI technologies to drive innovation and improve operational efficiencies; however, the successful adoption of AI hinges on an intricate interplay of organizational culture, user acceptance, strategic dimensions, and external environmental pressures.

First, organizational culture plays a pivotal role in shaping the approach a company takes towards AI adoption. Lee et al emphasize that the culture, values, and design of an organization significantly influence business model innovation, particularly in contexts where AI is integrated into service delivery [11]. Firms with cultures that prioritize innovation,

flexibility, and a willingness to embrace technology are better positioned to incorporate AI solutions effectively. It is vital for organizations to recognize that fostering a culture of openness and experimentation can facilitate the acceptance of AI initiatives among employees and encourage creative applications within various business processes.

Moreover, user trust and acceptance of AI technologies are critical determinants of successful integration. Malsia & Loku highlight that consumer attitudes towards AI devices significantly impact their acceptance, indicating that understanding these perceptions is essential for effective AI service implementation, especially in sensitive sectors such as health [2]. Enhancing user trust not only facilitates smoother implementation but also builds confidence in AI-driven solutions, prompting wider usage and engagement from both customers and employees. This understanding underscores that businesses must invest in education and transparency regarding AI functionalities to address concerns and foster positive user experiences.

Additionally, external market conditions and competitive dynamics influence the degree to which organizations prioritize AI in their strategies. Al-Matari et al discuss how dynamic capabilities, which encompass the organization's ability to adapt to changes in the external environment, play a significant role in shaping the integration of AI solutions [12]. Organizations that remain cognizant of shifts in technology, consumer behaviors, and competitive pressures are more adept at developing AI strategies that align with broader market trends. This adaptability not only fosters resilience but also empowers firms to leverage AI as a strategic asset in response to evolving industry dynamics.

Furthermore, analyzing the drivers and barriers to AI adoption across various sectors also reveals essential insights. Kar et al outline that organizations focusing on both the enabling factors and obstacles can tailor their strategies for greater effectiveness [13]. By understanding what hinders or facilitates AI integration—ranging from technical skills to management buy-in—businesses can develop comprehensive frameworks that support smooth transitions to AI-enhanced operations.

The primary objective of this research is to predict the percentage of AI service focus within companies, based on various company characteristics. By utilizing machine learning techniques such as Random Forest and Support Vector Machine, this study seeks to analyze how factors like company size, project scale, and industry domain influence the extent to which companies focus on AI services. Understanding these relationships can provide valuable insights into the drivers of AI adoption, assisting businesses in making data-driven decisions regarding their AI investments.

The significance of this study lies in its potential to help businesses better understand the key factors that drive AI adoption and integration within their operations. By predicting the level of AI service focus, companies can tailor their strategies, allocate resources more effectively, and enhance their competitiveness in an increasingly AI-driven market. The findings of this research will also contribute to the broader understanding of how organizational characteristics impact AI service focus, offering practical implications for businesses looking to optimize their AI initiatives.

2. Literature Review

2.1. AI Service Focus in Businesses

The integration of Artificial Intelligence (AI) into business strategies has garnered significant attention in scholarly research, demonstrating its transformative potential across various sectors. This review provides insights into key themes, including the role of organizational culture, strategic frameworks for adoption, and the dual effect of AI on business operations and relationships.

A foundational aspect of AI integration is the impact of organizational culture on the development of AI-driven business models. Lee et al. assert that organizational factors such as culture, design, and values significantly shape the innovation process related to AI, emphasizing the need for businesses to foster an environment conducive to experimentation and adaptability to leverage AI technologies effectively [11]. Companies that embrace change are better positioned to transition towards AI-driven models, enabling them to innovate and respond swiftly to evolving market demands.

The relationship between AI and human resources presents a nuanced dynamic, as AI can both enhance and disrupt business practices. Vorobeva et al. discuss how companies should prioritize human achievements to maintain customer

interest in service contexts, suggesting that while AI can enhance efficiency, a balance between automation and human interaction is necessary, especially in interpersonal service industries [14].

Moreover, exploring the drivers and barriers to AI adoption is crucial for understanding how firms can enhance their AI strategies. Kar et al. emphasize that recognizing these factors is vital for organizations to navigate the complexities of AI integration [13]. Understanding potential challenges—be they technological, infrastructural, or managerial—enables organizations to proactively address obstacles, facilitating smoother implementation processes and ensuring AI initiatives align with overall business objectives.

Lastly, the emerging discourse surrounding AI-driven business models underscores the necessity for companies to adapt their strategies continually. Farayola et al. explore how AI reshapes traditional business paradigms by examining the interaction between technological innovation and strategic management [7]. Their analysis underscores that businesses must remain vigilant regarding shifts in AI capabilities and continuously adapt their approaches to harness AI's full potential.

2.2. Machine Learning Algorithms for Prediction

The use of algorithms for predictive tasks has gained prominence in various fields, such as business, healthcare, and environmental sciences. Among the popular algorithms, Random Forest (RF) and Support Vector Machine (SVM) stand out for their effectiveness and adaptability in handling complex datasets.

Random Forest is an ensemble machine learning method that constructs a multitude of decision trees during training and outputs the mode of their predictions for classification tasks or the average for regression tasks. This technique is particularly beneficial because it reduces overfitting, a common issue with single decision trees, thus enhancing predictive accuracy. Ouyang demonstrated that RF outperformed both linear regression and decision tree models in predicting diamond prices due to its ability to mitigate the impact of individual tree outliers through ensemble learning [15]. Additionally, this method is characterized by its robustness and high tolerance for noise and outliers, as highlighted by Li et al., who noted that RF combines various models to yield better predictions compared to single models [16].

In practical applications, RF has been successfully employed in diverse domains such as aquaculture and healthcare. Fiesta's study illustrated how RF achieved 90% accuracy in predicting abiotic stress factors in aquaculture systems by leveraging multiple water quality indicators [17]. Furthermore, Iwendi et al. employed a boosted version of the RF algorithm to predict COVID-19 patient outcomes, demonstrating its efficacy in dealing with imbalanced datasets [18]. This versatility signifies RF’s potential across different contexts, suggesting that it is particularly well-suited for tasks with complex data interactions.

Support Vector Machine, on the other hand, is a supervised learning algorithm primarily used for classification tasks. It works by finding the hyperplane that best separates different classes in the feature space. SVM is particularly effective in high-dimensional spaces and is robust against overfitting, especially in cases where the number of dimensions exceeds the number of samples. Sheu demonstrated the applicability of classical SVM in classifying the priority of auditing XBRL instance documents, finding that SVM often outperforms other classification models such as logistic regression and decision trees [19]. This efficiency arises from its ability to create comprehensive classification margins that delineate between classes effectively.

In a comparative context with RF, Zhao et al. evaluated several models for predicting sorting center volume, including RF and SVM. Their findings revealed that while both algorithms yielded promising results, RF consistently demonstrated optimal performance across several evaluation metrics, including mean squared error (MSE) and root mean squared error (RMSE) [20]. This suggests that while SVM is a powerful algorithm in specific scenarios, RF may provide a more generalized solution for tasks requiring significant interpretive power across different data sets.

The choice between Random Forest and Support Vector Machine hinges on the specific predictive task and characteristics of the dataset involved. Random Forest offers exceptional performance in handling diverse, noisy datasets with its ensemble learning approach, making it highly popular in various applications, from environmental monitoring to customer satisfaction prediction [21]. Conversely, SVM excels in scenarios where a clear margin of separation between classes is paramount, particularly in high-dimensional settings.

2.3. Regression and Classification in Business Data

The application of regression models and classification techniques is essential in the field of data analysis, particularly in business contexts. By leveraging these analytical methods, organizations can derive actionable insights from complex datasets, enhance decision-making processes, and optimize operational outcomes. This discussion explores the significance of regression and classification models in business-related data analysis, drawing upon various research studies to highlight key aspects.

Regression analysis is employed to understand relationships between a dependent variable and one or more independent variables. It is particularly useful in predicting continuous outcomes, which is crucial for various business applications. For instance, Suwignjo et al. utilized regression analysis to forecast inventory performance, categorizing products into understock, normal, and overstock classes based on predictive analytics. Their study demonstrated that regression models could provide significant insights into inventory levels and help businesses optimize stock management [22].

Implementing regression methods, particularly Generalized Linear Models (GLMs), allows for effective modeling of diverse business scenarios. GLMs, as characterized by Jas et al., offer advantages such as interpretability and the ability to accommodate various distribution types. This enhances their utility in real-world applications across sectors like marketing analysis and sales forecasting, facilitating better strategic planning and resource allocation [23].

Classification techniques are essential for categorizing data into distinct classes based on input features. These models are frequently applied in scenarios where the outcome variable is categorical. For example, in predicting the productivity of manufacturing machines, Sinlae et al. compared Naïve Bayes and decision tree classification algorithms, illustrating how different classification approaches can impact predictive accuracy. Their findings underscore the importance of selecting the appropriate classification algorithm based on the specific context and data characteristics [24].

Moreover, the convergence of these analytical techniques with business intelligence (BI) enhances the ability of organizations to convert raw data into strategic insights. As described by Elseddawy and Hegazy, BI incorporates a wide array of data science techniques, including regression and classification, to distill meaningful information from extensive and varied datasets. This ability to analyze multifaceted data allows businesses to develop actionable strategies across various operations, such as consumer behavior analysis, sales forecasting, and inventory management [25].

3. Method

3.1. Data Collection and Preprocessing

The dataset used in this study is collected from various AI companies, containing critical information on company characteristics and their AI service focus. The first step involves loading the dataset from a CSV file using the `pandas` library. Upon loading, the data undergoes a cleaning process to ensure that only relevant and complete entries are retained for analysis. Initially, rows with missing values in the 'Website' column are removed, as the absence of a valid website often indicates that the data is incomplete or irrelevant for the analysis. Additionally, identifying columns such as 'Company_Name' and 'Website' are dropped, as they do not contribute directly to the predictive model. These columns primarily serve as identifiers but do not provide useful information for predicting the target variable, which is the percentage of AI service focus.

A more thorough cleaning process is then applied to the specific columns that contain critical information for our model. The 'Percent AI Service Focus' column is transformed by removing the percentage sign and converting its values into numeric format. Any non-numeric entries that cannot be converted are either dropped or imputed, ensuring that only valid data remains for analysis. Similarly, the columns 'Minimum Project Size', 'Average Hourly Rate', and 'Number of Employees' require custom functions to clean their respective values. The 'Minimum Project Size' and 'Average Hourly Rate' columns contain ranges or textual representations (such as 'undisclosed' or '\$50 - \$100'), so these are cleaned by extracting the numerical values and converting them into a consistent numeric format. Any missing values or unclear entries, like 'undisclosed' or text without numerical data, are assigned NaN (Not a Number), which will later be imputed or dropped, ensuring that only usable data is included in the analysis. Similarly, the 'Number of

Employees' column is cleaned by converting entries like "freelancer" to 1, or by extracting numeric values for ranges, such as "250 - 999", by taking the average of the range. This entire cleaning procedure ensures that all columns are standardized and ready for the machine learning models.

3.2. Exploratory Data Analysis (EDA)

Prior to building predictive models, an extensive exploratory data analysis (EDA) is conducted to gain a deeper understanding of the dataset and the relationships between different features. The first step in the EDA process is to examine the distribution of the target variable, 'Percent AI Service Focus', which is the key outcome we aim to predict. A histogram of this variable is plotted to visualize its distribution, and a kernel density estimate (KDE) is overlaid to understand the underlying distribution better. This gives insights into whether the data is normally distributed or skewed, which could inform the choice of modeling techniques. Following this, an analysis of the numerical features is conducted by plotting their distributions through scatter plots, identifying any outliers or anomalies. Additionally, basic statistics, such as mean, median, and standard deviation, are calculated for the numerical columns like 'Minimum Project Size', 'Average Hourly Rate', and 'Number of Employees'. The relationships between these features and the target variable are also explored. For example, scatter plots and box plots are used to investigate how 'Average Hourly Rate' and 'Minimum Project Size' relate to 'Percent AI Service Focus', providing insights into whether these company characteristics have any significant influence on the focus of AI services.

In addition to numerical features, categorical variables like 'Location' are also examined. The distribution of locations is visualized using a count plot, which shows the frequency of companies operating in different geographic locations. A correlation matrix is also calculated to identify any potential linear relationships between the numerical features and the target variable. This helps to uncover any obvious correlations, such as whether companies with a larger number of employees or higher average hourly rates tend to have a higher AI service focus. Through these visualizations and statistical measures, the EDA phase helps inform the next steps in modeling, providing a clearer picture of the data's structure and potential relationships.

3.3. Modeling

The primary objective of this research is to predict the percentage of AI service focus within companies based on their characteristics. To achieve this, two machine learning algorithms are employed: Random Forest Regressor and Support Vector Regressor (SVR). Both models are well-suited for regression tasks and can handle complex, non-linear relationships within the data. The Random Forest algorithm is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and prevent overfitting. The SVR, on the other hand, is particularly useful for handling data with high variance and is known for its effectiveness in capturing the underlying patterns in the dataset.

To prepare the data for modeling, the dataset is first split into training and testing sets using an 80/20 split, ensuring that the model is trained on a large portion of the data while still being evaluated on a separate testing set to assess its generalization performance. Prior to training the models, a preprocessing pipeline is set up using the `ColumnTransformer` from the `sklearn` library. This pipeline handles the transformation of different feature types in the dataset. Numerical features are first imputed using the median value to handle missing data, and then standardized using the `StandardScaler`. Categorical features are imputed with the value 'Unknown' and encoded using one-hot encoding to convert them into a format suitable for machine learning algorithms. This ensures that both numerical and categorical features are appropriately preprocessed before being fed into the models.

Once the preprocessing is complete, both the Random Forest and SVR models are trained on the training set. The Random Forest model is configured with 100 estimators and parameters like `max_depth=10` and `min_samples_split=10` to prevent overfitting. The SVR model is configured with default parameters, but its performance is highly dependent on the tuning of hyperparameters like `C`, `gamma`, and `epsilon`. After training, both models are evaluated on the test set to assess their predictive performance. The evaluation metrics used include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2), which provide insights into the models' accuracy and their ability to predict the percentage of AI service focus.

3.4. Model Evaluation

After training the models, they are evaluated based on their performance on the test set. The evaluation process involves comparing the predicted values with the actual values of 'Percent AI Service Focus' to calculate the error metrics. MAE is used to measure the average magnitude of the errors in the predictions, MSE and RMSE provide insights into the variance of the errors, and R^2 quantifies the proportion of variance in the target variable that is explained by the models. These metrics allow for a comparison between the Random Forest and SVR models, identifying which model performs better in predicting the AI service focus. Visualizations, such as scatter plots of the actual vs. predicted values, are also used to provide a visual representation of how well the models are performing.

Additionally, the feature importance from the Random Forest model is extracted to understand which company characteristics have the greatest influence on the AI service focus prediction. This is particularly valuable for businesses looking to understand the key factors that determine AI adoption and how to focus their resources accordingly. Feature importance scores are plotted to visualize the relative contributions of each feature, helping to interpret the model and provide actionable insights for businesses.

3.5. Feature Engineering and Final Model

After evaluating both models, the study proceeds with a deeper analysis of the feature importance in the Random Forest model. This helps to identify which company characteristics, such as project size, hourly rate, number of employees, and location, are most influential in predicting the AI service focus. Based on this, further feature selection can be performed to refine the model and improve its accuracy. In addition, hyperparameter tuning may be performed on the SVR model to enhance its performance. Finally, the models are fine-tuned, and the best-performing model is selected for making predictions on new data.

The final model is then used to provide insights into the key factors influencing AI adoption within companies. By understanding these factors, businesses can optimize their strategies for AI service implementation and better allocate resources for AI-related projects. The model's predictions can also assist companies in making informed decisions about how much to invest in AI services based on their characteristics.

4. Results and Discussion

4.1. Results of Data Cleaning and Initial Observations

The dataset initially contained 3100 rows and 8 columns. After loading the data, the first few rows revealed that columns such as 'Company_Name', 'Website', and 'Unnamed: 7' were not useful for the prediction task and were therefore removed. Additionally, 1 row with a missing 'Website' value was dropped, reducing the dataset to 3099 rows and 7 columns. The remaining columns, including 'Location', 'Minimum Project Size', 'Average Hourly Rate', 'Number of Employees', and 'Percent AI Service Focus', were retained for analysis.

Several data cleaning operations were performed on specific columns to ensure consistency and compatibility with machine learning models. The 'Percent AI Service Focus' column, originally in percentage format, was cleaned by removing the '%' sign and converting the values to numeric format. This new column, 'Target_Percent_AI', successfully represented the AI service focus as numerical values (e.g., 10%, 15%, and 40%). The 'Minimum Project Size' column, which had text entries like "Undisclosed" and ranges like "\$1,000+", was cleaned by extracting numerical values where possible, resulting in a new column, 'Clean_Min_Project_Size', with numeric representations such as 1000, 10000, and 50000.

Similarly, the 'Average Hourly Rate' column was cleaned by converting ranges like "\$50 - \$99 / hr" and entries like "< \$25 / hr" into numeric values, resulting in 'Clean_Avg_Hourly_Rate'. For the 'Number of Employees' column, ranges such as "250 - 999" were converted to their midpoints, producing the 'Clean_Num_Employees' column. The 'Location' column was simplified by extracting the state or country abbreviation and grouping less frequent locations into a category called 'Other'. This resulted in the 'Clean_Location' column, where locations such as "Los Altos, CA" were standardized to "CA" and "Portland, OR" was categorized as "Other".

After the initial cleaning process, missing values were identified in several columns. The 'Clean_Min_Project_Size' column had 1158 missing values, the 'Clean_Avg_Hourly_Rate' column had 1247 missing values, while 'Clean_Num_Employees' and 'Clean_Location' had no missing values. Importantly, the target variable, 'Target_Percent_AI', had no missing values, making it ready for model training. The missing values in 'Clean_Min_Project_Size' and 'Clean_Avg_Hourly_Rate' are significant, as these features are crucial for predicting AI service focus. These missing values will be handled during the model preparation phase, where imputation techniques such as median imputation will be applied to fill in the gaps.

4.2. Exploratory Data Analysis (EDA)

The EDA phase revealed key insights into the distribution of the data. The target variable, 'Target_Percent_AI', showed a relatively balanced distribution, with companies having AI service focus percentages ranging from 10% to 40%. Histograms and kernel density plots were used to visualize the distribution of this target variable, helping to identify any skewness or outliers in the data.

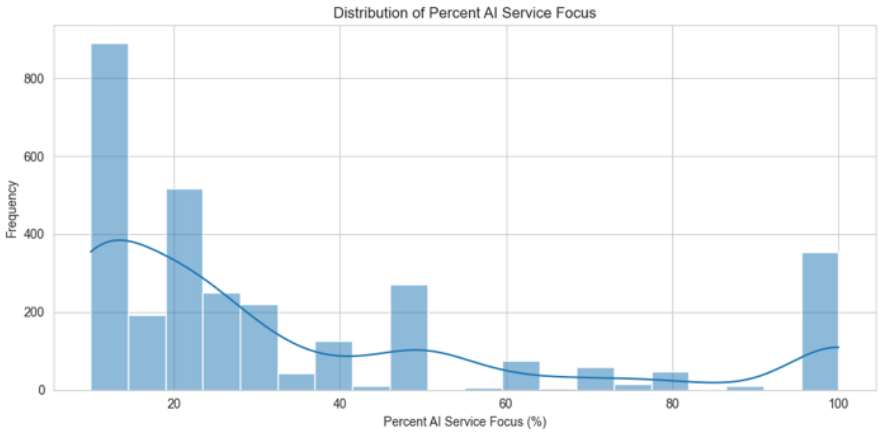

Figure 1. Distribution of Percent AI Service Focus

Figure 1 displays the distribution of the target variable, 'Percent AI Service Focus'. The histogram shows that most companies in the dataset have a low focus on AI, with a significant portion (around 800 companies) having only 10% focus. The distribution is skewed, with a small number of companies exhibiting high AI service focus, especially around 100%. The smooth curve (KDE) reveals that, beyond the initial peak, the distribution declines gradually, indicating fewer companies with higher AI service focus. This suggests that AI adoption is still relatively limited among the majority of companies, with few leading companies having a large focus on AI services.

For the numerical features, such as 'Clean_Min_Project_Size', 'Clean_Avg_Hourly_Rate', and 'Clean_Num_Employees', descriptive statistics were calculated, and visualizations were generated to examine the spread and any potential outliers in the data. The 'Clean_Avg_Hourly_Rate' and 'Clean_Min_Project_Size' columns had clear distributions, while 'Clean_Num_Employees' exhibited a more concentrated distribution around the mid-range values, indicating a preference for companies with 50 to 250 employees.

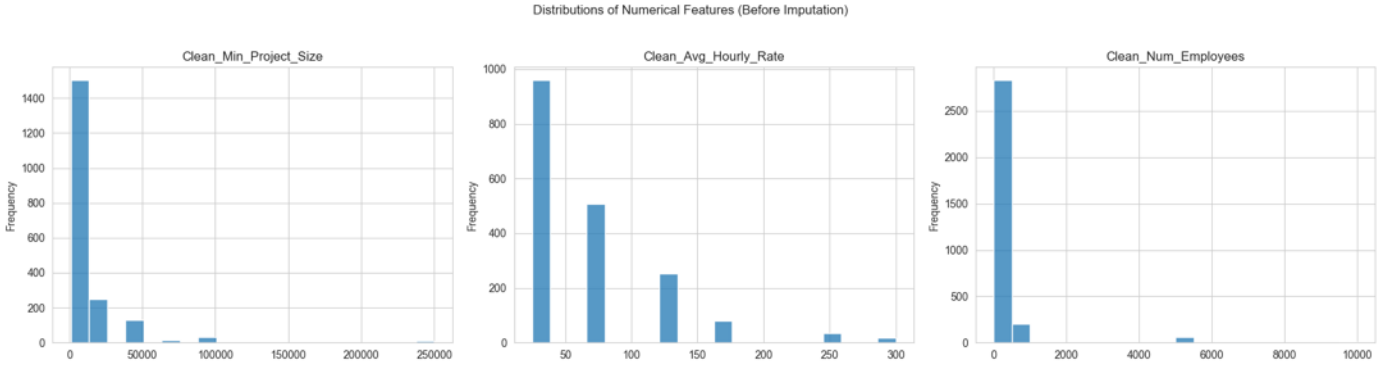

Figure 2. Distribution of Numerical Features

 ISSN 2776-8007

Figure 2 displays the distributions of the numerical features before imputation. Clean_Min_Project_Size exhibits a highly skewed distribution, with the majority of companies having smaller projects (close to 0), but a small number of companies handle much larger projects (up to 250,000). The distribution reflects that most companies in the dataset deal with smaller or medium-scale projects, with very few involved in large-scale projects. Clean_Avg_Hourly_Rate also follows a skewed distribution, with the majority of companies charging an hourly rate between $50 and $100, indicating a concentration of businesses offering moderately priced services. As the hourly rate increases, the frequency decreases, with a few companies charging significantly higher rates, but those are outliers. Clean_Num_Employees shows a high concentration of companies with fewer employees, especially those in the range of 50-250 employees, reflecting the prevalence of small to mid-sized companies in the dataset. The distribution tapers off as the number of employees increases, with only a small number of companies having large workforces (up to 6,000 employees).

Categorical data, particularly 'Clean_Location', was also analyzed. Locations like "CA" and "NY" had the most significant representation in the dataset, while other regions were grouped into the "Other" category to simplify the analysis and prevent overfitting from too many categories. The EDA revealed that the majority of companies in the dataset were located in a few prominent regions, with a small number of companies distributed across other locations.

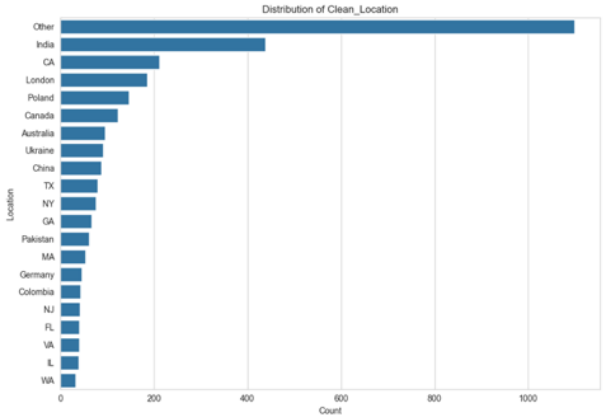

Figure 3. Distribution of Clean_Location

Figure 3 visualizes the distribution of the 'Clean_Location' feature, showing the frequency of companies across various geographic locations. The bar chart indicates that the majority of companies in the dataset are either located in India or California (CA), which have the highest frequencies, followed by London and Poland. A significant portion of the data is also grouped under the "Other" category, suggesting that companies are spread across numerous other locations globally, but in smaller numbers. The chart illustrates the diversity of company locations, with multiple countries and regions represented. However, the dominance of a few regions, particularly India and CA, points to the concentration of companies in certain geographic areas, likely influenced by regional markets and access to resources or talent pools. This geographical information could be useful for understanding trends in AI service adoption based on location.

 ISSN 2776-8007

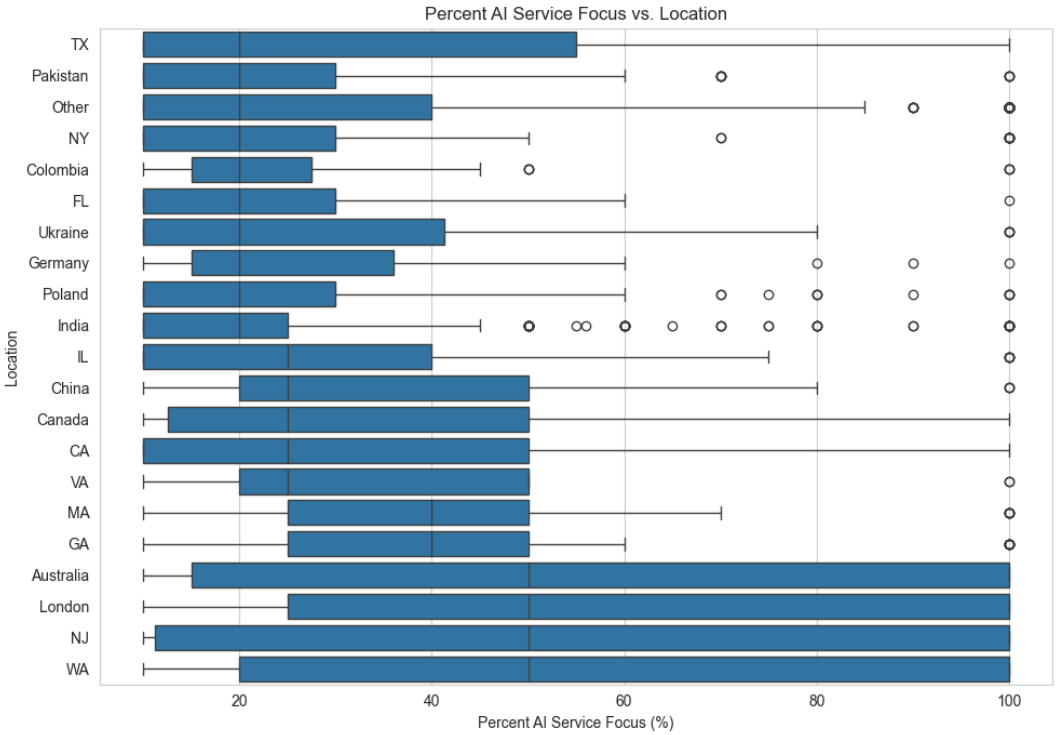

Figure 4. Percent AI Service Focus vs. Location

Figure 4 is a boxplot showing the distribution of AI service focus percentages across different locations. Each location is represented by a horizontal boxplot, where the central mark indicates the median value, the box represents the interquartile range (IQR), and the whiskers show the spread of the data. Several locations, such as Australia and London, have a broader range of AI service focus, indicating that companies in these areas vary widely in their commitment to AI services. In contrast, locations like TX, Pakistan, and Other show a relatively more concentrated focus, with the boxplots indicating a smaller spread in AI service focus values. Outliers are visible in some locations, particularly China, Canada, and Germany, where a few companies report very high AI service focus, likely indicating companies that are heavily invested in AI initiatives. On the other hand, locations like India and Poland have a wider range of AI focus percentages, with some companies showing minimal to moderate AI adoption, while others are more focused. The plot indicates that certain regions may be more consistent in their AI adoption, while others show a significant spread in AI focus, reflecting the diversity in business models and technological investment across different countries. This figure provides insights into how geographical factors may influence AI service adoption, with some countries exhibiting a stronger focus on AI compared to others. It can be used to identify regions where AI adoption is either widespread or in its early stages.

The correlation matrix was also computed to understand the relationships between the numerical features and the target variable. This analysis showed that 'Clean_Min_Project_Size' and 'Clean_Avg_Hourly_Rate' had moderate correlations with 'Target_Percent_AI', indicating that larger projects and higher hourly rates are likely associated with a higher focus on AI services. However, 'Clean_Num_Employees' and 'Clean_Location' showed weaker correlations with the target variable, suggesting that company size and location may have a less direct impact on AI service focus compared to project size and hourly rates.

 ISSN 2776-8007

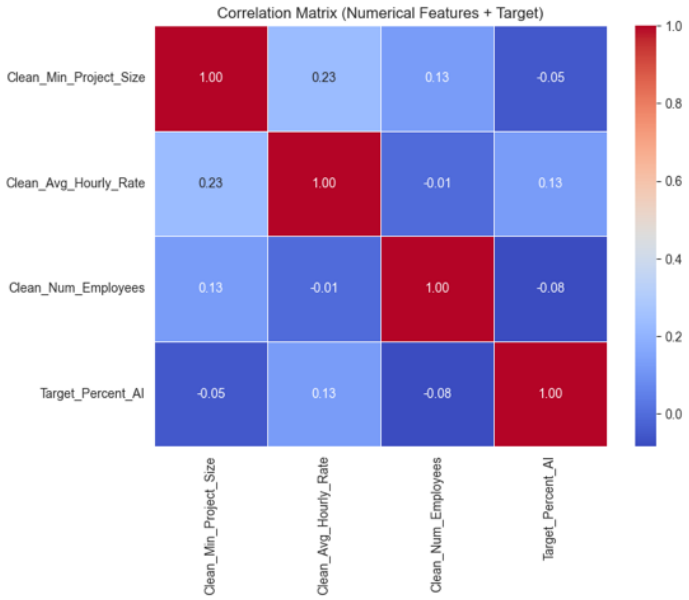

Figure 5. Correlation Matrix (Numerical Features + Target)

Figure 5 shows a correlation matrix that evaluates the relationships between the numerical features and the target variable, 'Target_Percent_AI'. The matrix provides a clear visual representation of how each feature correlates with others and the target variable. Clean_Min_Project_Size has a weak positive correlation with Clean_Avg_Hourly_Rate (0.23) and Clean_Num_Employees (0.13). This indicates that companies with larger projects tend to have slightly higher hourly rates and a moderate number of employees. However, the correlation is relatively low, suggesting that the relationship between project size, hourly rates, and employee numbers is not very strong. The Target_Percent_AI variable shows a negative correlation with Clean_Min_Project_Size (-0.05), which suggests that there is little to no direct relationship between the minimum project size and the focus on AI services. Clean_Avg_Hourly_Rate correlation with Clean_Min_Project_Size (0.23) is positive, indicating that companies charging higher hourly rates tend to have slightly larger projects. Interestingly, Clean_Avg_Hourly_Rate has almost no correlation with Clean_Num_Employees (-0.01), suggesting that a company's hourly rate does not necessarily depend on its number of employees. The Target_Percent_AI variable has a positive correlation with Clean_Avg_Hourly_Rate (0.13), indicating that companies charging higher hourly rates might have a slightly higher focus on AI services, but again, the correlation is weak. Clean_Num_Employees exhibits very weak correlations with both Clean_Min_Project_Size (0.13) and Clean_Avg_Hourly_Rate (-0.01), suggesting that the number of employees does not have a strong influence on the size of the projects or the hourly rates. Additionally, the Clean_Num_Employees feature has a very weak negative correlation with Target_Percent_AI (-0.08), indicating that the number of employees in a company does not strongly affect its focus on AI services. Target_Percent_AI shows very weak correlations with all the numerical features, including Clean_Min_Project_Size (-0.05), Clean_Avg_Hourly_Rate (0.13), and Clean_Num_Employees (-0.08). These weak correlations suggest that while company characteristics like project size, hourly rate, and employee count have some influence on AI service focus, they are not the most significant predictors. This highlights the need for additional features or more advanced modeling techniques to improve the accuracy of AI focus predictions.

The target variable, 'Target_Percent_AI', represents the percentage of AI service focus within companies. This variable exhibited a range from 10% to 100%, with the majority of companies concentrated in the lower and middle ranges of AI adoption. The mean percentage of AI service focus across all companies is approximately 34.37%, with a standard deviation of 29.1%, indicating a relatively high variability in AI adoption among the companies in the dataset. The median value of 20% suggests that most companies have a lower focus on AI services, with the interquartile range (IQR) spanning from 10% (25th percentile) to 50% (75th percentile). These statistics indicate that while there are a few companies with very high AI focus (up to 100%), the majority have a relatively low to moderate focus on AI, highlighting potential opportunities for AI growth in the industry.

The statistics for the numerical features—'Clean_Min_Project_Size', 'Clean_Avg_Hourly_Rate', and 'Clean_Num_Employees'—reveal significant insights into the dataset's structure and the nature of the companies represented. The minimum project size with values ranging from 1,000 to 250,000. The mean project size is 14,746, but there is substantial variability, with a standard deviation of 25,891. The 25th percentile indicates that 25% of companies work on projects worth at least 5,000, while the 75th percentile shows that 75% of companies handle projects worth at least 10,000. The high maximum value of 250,000 suggests the presence of outliers or companies involved in very large-scale projects. The average hourly rate ranges from 25 to 300, with a mean value of 70.06 and a standard deviation of 52.06. This indicates a wide disparity in hourly rates charged by companies, likely reflecting differences in the types of services offered and the expertise required. The 25th and 75th percentiles show that the majority of companies have hourly rates between 37 and 74.5, with the highest values reaching up to 300 per hour. This wide range in hourly rates likely corresponds to the varying levels of service specialization and the industries in which the companies operate. The number of employees in the companies ranges from as few as 1 to as many as 10,000. The mean number of employees is 228.86, but with a high standard deviation of 943.62, suggesting that the dataset includes both small startups and large organizations. The 25th percentile value of 29.5 employees and the 75th percentile value of 149.5 employees indicate that most companies are small to medium-sized, while a small number of companies have a much larger workforce, skewing the maximum value to 10,000.

The correlation between numerical features and the target variable, 'Target_Percent_AI', is also worth noting. The project's minimum size and hourly rate showed moderate correlations with the target variable, indicating that companies with larger projects and higher hourly rates tend to focus more on AI services. This is consistent with the intuition that larger, more resource-intensive projects would necessitate a higher degree of AI integration. The number of employees showed weaker correlations with AI focus, suggesting that company size may not be as strongly related to AI adoption as other factors like project size and service pricing. These insights, coupled with the statistical analysis, help in understanding which company characteristics are most influential in determining the level of AI service focus, guiding the feature selection and model training in the subsequent stages.

4.3. Model Evaluation Results

The models were trained using two different machine learning algorithms: Random Forest Regressor and Support Vector Regressor (SVR). The training process was efficient, with the Random Forest model completing training in just 0.21 seconds, while the SVR model took 0.28 seconds to train. The Random Forest algorithm is an ensemble method that combines multiple decision trees to provide a more robust prediction, while the SVR algorithm is designed to capture non-linear relationships in the data. Both models were trained on the same training set, with their performance evaluated based on their ability to predict the percentage of AI service focus (the target variable) in companies. The performance of both models was assessed using several key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). The results for the Random Forest Regressor were as follows: MAE: 20.43, MSE: 723.64, RMSE: 26.90 and R^2: 0.12. The Support Vector Regressor (SVR) showed slightly better MAE and MSE, with MAE: 19.42, MSE: 849.97, RMSE: 29.15 andR^2: -0.03. From these results, we can observe that the Random Forest model has a slightly higher R^2 value, indicating that it is a better fit for the data in terms of explaining variance in the target variable. On the other hand, the SVR model's negative R^2 suggests that it performs poorly on this dataset, as it is not able to model the relationship between the features and target effectively. The relatively low R^2 values for both models indicate that there are other factors affecting AI service focus that are not captured by the available features.

 ISSN 2776-8007

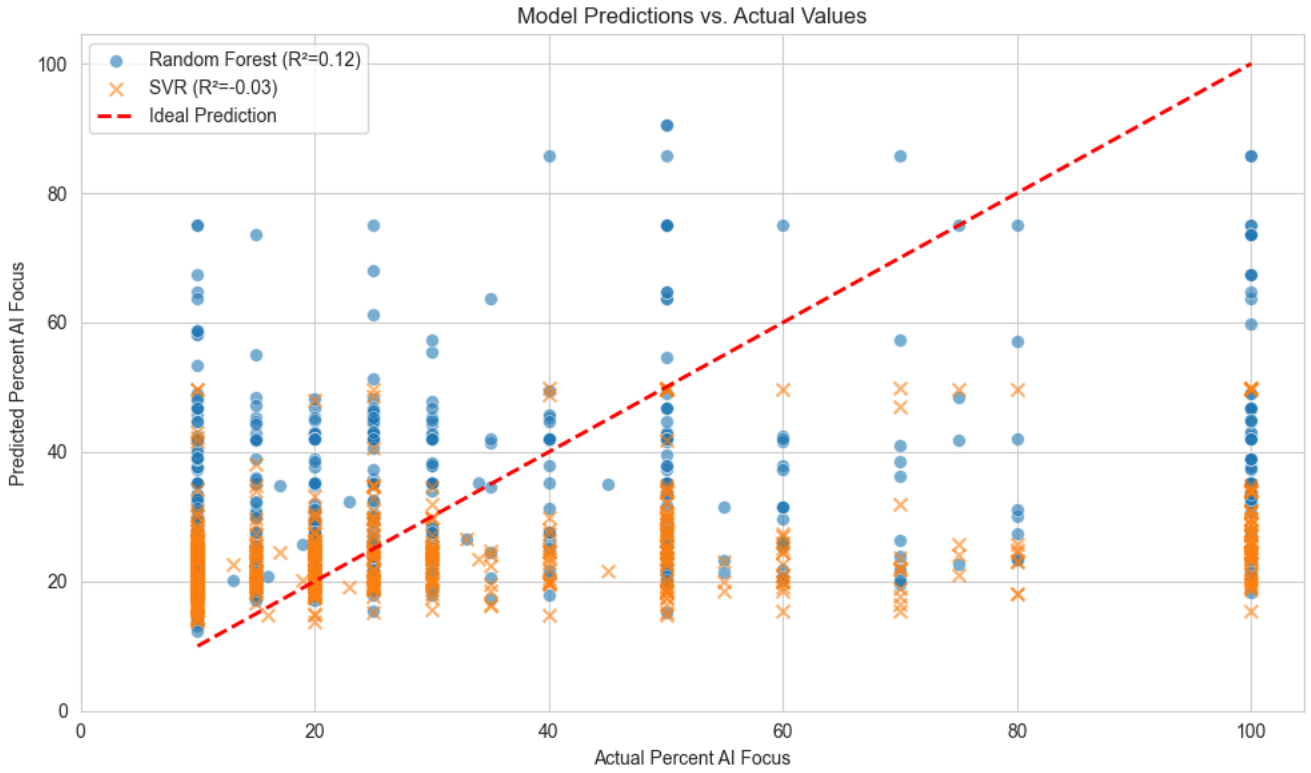

Figure 6. Model Predictions vs. Actual Values

Figure 6 presents a scatter plot comparing the predicted and actual values of Percent AI Service Focus for both the Random Forest and Support Vector Regressor (SVR) models. The Random Forest model is represented by blue circles, and the SVR model is shown as orange crosses. The ideal prediction line (dashed red) represents the scenario where the predicted values perfectly match the actual values. From the plot, we can observe that the Random Forest model has a moderate fit to the data, as indicated by the points clustering closer to the ideal prediction line, especially for lower AI service focus percentages. However, there are still noticeable deviations, particularly for higher focus percentages, where the predictions tend to be less accurate. On the other hand, the SVR model shows a wider spread of predicted values, with many points deviating significantly from the ideal line. This suggests that the SVR model performed poorly, as its predictions are less aligned with the actual values, especially in the higher AI focus range. Overall, the Random Forest model appears to provide a better fit compared to the SVR model, as it shows a higher concentration of points near the ideal line, though improvements could still be made. The scatter plot visually emphasizes the models' predictive capabilities and highlights areas where further refinement or more features may be necessary to improve the accuracy of AI service focus predictions.

4.4. Feature Importance Analysis

The feature importance scores for the Random Forest Regressor model were calculated to determine which features contributed most to the prediction of AI service focus. These results show that the Clean_Min_Project_Size and Clean_Avg_Hourly_Rate features were the most important predictors of AI service focus, with Clean_Location_London and Clean_Num_Employees also playing significant roles. This suggests that companies with larger projects and higher hourly rates are more likely to have a higher focus on AI services. The location-related features, such as Clean_Location_London and Clean_Location_Australia, indicate that certain geographical regions might be associated with more intense AI adoption, which could reflect regional market trends or the availability of AI talent. The visualization of feature importance provides a clear understanding of how each feature contributes to the predictions, which can be valuable for businesses to focus on the most influential factors when strategizing AI adoption. However, the relatively low R^2 values indicate that more features or a different modeling approach may be required to achieve a more accurate prediction of AI service focus across companies.

 ISSN 2776-8007

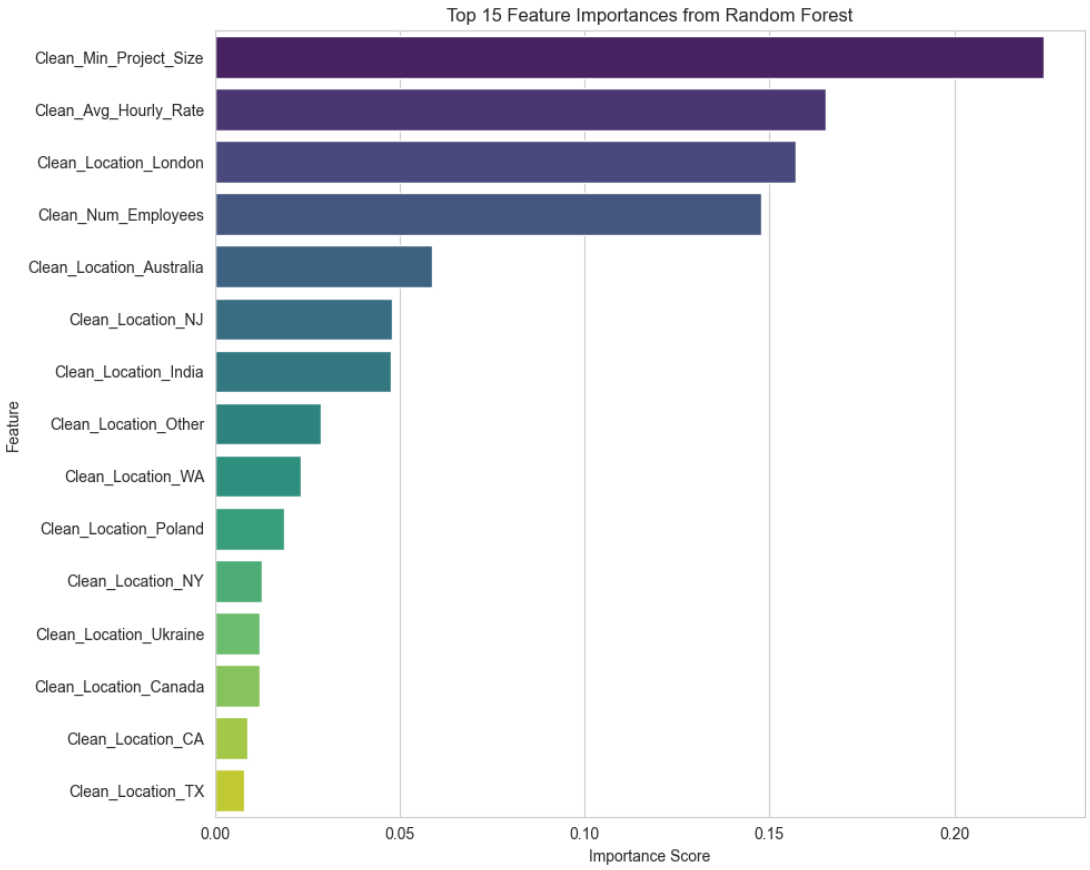

Figure 7. Top 15 Feature Importances from Random Forest

Figure 7 visualizes the Top 15 Feature Importances derived from the Random Forest model. The chart highlights the significance of each feature in predicting the 'Target_Percent_AI' variable, which represents the AI service focus in companies. The most important feature is Clean_Min_Project_Size, with the highest importance score, indicating that companies with larger project sizes are more likely to have a higher focus on AI services. Following this, Clean_Avg_Hourly_Rate and Clean_Location_London also appear as significant predictors, showing that companies with higher hourly rates or those located in London are associated with a stronger AI service focus. Clean_Num_Employees is also an important feature, though its influence is relatively lower compared to project size and hourly rate. Geographical locations, such as Clean_Location_Australia, Clean_Location_NJ, and Clean_Location_India, also contribute to the model, with varying degrees of importance. The Clean_Location_Other category, which groups less frequent locations, is also featured, suggesting that some regions not specifically listed still play a role in AI service focus. These results reflect how both project-related characteristics and regional factors influence the level of AI adoption among companies.

5. Conclusion

This study aimed to predict the level of AI service focus in companies using machine learning models, specifically the Random Forest Regressor and Support Vector Regressor (SVR). The Random Forest model performed slightly better than the SVR, with an R^2 value of 0.12, indicating a modest ability to explain the variance in AI service focus. The SVR model, however, had a negative R^2 value (-0.03), suggesting it was less effective in modeling the relationship between company characteristics and AI adoption. The analysis revealed that factors such as Clean_Min_Project_Size and Clean_Avg_Hourly_Rate were significant predictors of AI service focus, with the Random Forest model providing valuable insights into the importance of these features. Despite the promising results, the study has several limitations. The models exhibited relatively low predictive performance, as evidenced by the R^2 values. This suggests that the

available features, such as company size, project size, and hourly rates, only partially explain the variance in AI service focus. Additionally, there were significant missing values in key columns, such as Clean_Min_Project_Size and Clean_Avg_Hourly_Rate, which were imputed but may have impacted the models' accuracy. Furthermore, the dataset lacks information on other potential predictors of AI adoption, such as the company's technological infrastructure, budget allocation for AI, or industry-specific trends, which could have improved the models' performance. To improve predictions of AI service focus, future research could incorporate additional business features that directly impact AI adoption, such as a company's R\&D expenditure, technological readiness, or market competition. Including more granular data on AI projects, such as the type of AI technologies implemented or the duration of AI projects, could provide deeper insights into the factors influencing AI focus. Additionally, advanced machine learning techniques, such as deep learning models or ensemble methods, could be explored to capture complex patterns in the data. Hyperparameter tuning and cross-validation could also be employed to improve the models' accuracy and robustness. Finally, expanding the dataset to include more diverse companies from different industries and regions could further enhance the generalizability of the models and their applicability to a wider range of businesses.

6. Declarations

6.1. Author Contributions

Conceptualization: T.S., L.T., and T.P.; Methodology: L.T.; Software: T.S.; Validation: T.S., L.T., and T.P.; Formal Analysis: T.S., L.T., and T.P.; Investigation: T.S.; Resources: L.T.; Data Curation: L.T.; Writing—Original Draft Preparation: T.S., L.T., and T.P.; Writing—Review and Editing: L.T., T.S., and T.P.; Visualization: T.S. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Hofacker, Book Review: Artificial Intelligence for Sustainable Value Creation, Management and Business Review, vol. 3, no. 1, 2023. doi: 10.1177/2694105820230301011.

[2] A. Zhukovska, T. Zheliuk, D. Shushpanov, V. Brych, O. Brechko, and N. Kryvokulska, “Management of the Development of Artificial Intelligence in Healthcare,” in *Proc. 2023 13th Int. Conf. on Advanced Computer Information Technologies (ACIT)*, pp. 241–247, 2023. doi: 10.1109/ACIT58437.2023.10275435.

[3] M. Huang and R. T. Rust, “Artificial Intelligence in Service,” *J. Serv. Res.*, 2018, doi: 10.1177/1094670517752459.

[4] J. Chin, C. Do, and M. Kim, “How to Increase Sport Facility Users’ Intention to Use AI Fitness Services: Based on the Technology Adoption Model,” *Int. J. Environ. Res. Public. Health*, 2022, doi: 10.3390/ijerph192114453.

[5] A. Fiske, P. Henningsen, and A. Buyx, “Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy,” *J. Med. Internet Res.*, 2019, doi: 10.2196/13216.

[6] J. A. Vilaginés, “Effective Integration of Artificial Intelligence: Key Axes for Business Strategy,” *J. Bus. Strategy*, 2023, doi: 10.1108/jbs-01-2023-0005.

[7] O. A. Farayola, A. A. Abdul, B. O. Irabor, and E. C. Okeleke, "Innovative Business Models Driven by Ai Technologies: A Review," *Comput. Sci. It Res. J.*, 2023, doi: 10.51594/csitrj.v4i2.608.

[8] A. K. Tiwari and D. Saxena, "Application of Artificial Intelligence in Indian Banks," in *Proc. 2021 Int. Conf. on Computational Performance Evaluation (ComPE)*, pp. 545–548, 2021. doi: 10.1109/ComPE53109.2021.9751981.

[9] T. Carpenter, "Revolutionising the Consumer Banking Experience with Artificial Intelligence," *Journal of Digital Banking*, vol. 5, no. 3, 2020.

[10] P. Agarwal, "Redefining Banking and Financial Industry Through the Application of Computational Intelligence," in *2019 Advances in Science and Engineering Technology International Conferences (ASET)*, pp. 1–5, 2019. doi: 10.1109/ICASET.2019.8714305.

[11] J. Lee, T. Suh, D. Roy, and M. S. Baucus, "Emerging Technology and Business Model Innovation: The Case of Artificial Intelligence," *J. Open Innov. Technol. Mark. Complex.*, 2019, doi: 10.3390/joitmc5030044.

[12] A. S. Al-Matari, R. Amiruddin, K. A. Aziz, and M. A. Al-Sharafi, "The Impact of Dynamic Accounting Information System on Organizational Resilience: The Mediating Role of Business Processes Capabilities," *Sustainability*, 2022, doi: 10.3390/su14094967.

[13] S. Kar, A. K. Kar, and M. P. Gupta, "Modeling Drivers and Barriers of Artificial Intelligence Adoption: Insights From a Strategic Management Perspective," *Intell. Sys Acc.*, 2021, doi: 10.1002/isaf.1503.

[14] W. Verleyen and W. McGinnis, "Framework for Disruptive AI/ML Innovation," *arXiv preprint*, arXiv:2204.12641, 2022. doi: 10.48550/arXiv.2204.12641.

[15] H. Zhang, "Prediction and Feature Importance Analysis for Diamond Price Based on Machine Learning Models," *Adv. Econ. Manag. Polit. Sci.*, 2023. doi: 10.54254/2754-1169/46/20230347.

[16] X. Li, J.-L. Wang, Z. Geng, Y. Jin, and J. Xu, "Short-Term Wind Power Prediction Method Based on Genetic Algorithm Optimized XGBoost Regression Model," *J. Phys. Conf. Ser.*, 2023, doi: 10.1088/1742-6596/2527/1/012061.

[17] T. Li, J. Lu, J. Wu, Z. Zhang, and L. Chen, "Predicting Aquaculture Water Quality Using Machine Learning Approaches," *Water*, vol. 14, no. 18, 2022. doi: 10.3390/w14182836.

[18] C. Iwendi *et al.*, "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm," *Front. Public Health*, 2020, doi: 10.3389/fpubh.2020.00357.

[19] G.-Y. Sheu, "Classification of the Priority of Auditing XBRL Instance Documents With Fuzzy Support Vector Machines Algorithm," *J. Auton. Intell.*, 2019, doi: 10.32629/jai.v2i2.40.

[20] Q. Ma, "Automatic Pricing and Replenishment Decision of Vegetable Commodities Based on Random Forest and ARIMA Models," *in Proc. 2023 IEEE Int. Conf. on Electrical, Automation and Computer Engineering (ICEACE)*, pp. 1496–1501, 2023. doi: 10.1109/ICEACE60673.2023.10441901.

[21] R. Luo, "Improved Random Forest Based on Grid Search for Customer Satisfaction Prediction," *Adv. Econ. Manag. Polit. Sci.*, 2023, doi: 10.54254/2754-1169/38/20231913.

[22] P. Suwignjo, L. Panjaitan, A. R. Baihaqy, and A. Rusdiansyah, "Predictive Analytics to Improve Inventory Performance: A Case Study of an FMCG Company," *Oper. Supply Chain Manag. Int. J.*, 2023, doi: 10.31387/oscm0530390.

[23] M. Jas *et al.*, "Pyglmnet: Python Implementation of Elastic-Net Regularized Generalized Linear Models," *J. Open Source Softw.*, 2020, doi: 10.21105/joss.01959.

[24] F. Sinlae, A. S. Yudhasti, and A. Wibowo, "Comparative Analysis of Naïve Bayes and Decision Tree Algorithms in Data Mining Classification to Predict Weckerle Machine Productivity," *J. Syst. Eng. Inf. Technol. Joseit*, 2022, doi: 10.29207/joseit.v1i2.3439.

[25] A. I. B. ElSeddawy and M. Hegazy, "A Proposed Mining Approach Based on Business Intelligence by Using Data Sciences Techniques," *Int. J. Intell. Comput. Inf. Sci.*, 2021, doi: 10.21608/ijicis.2021.57633.1052.