Umut Er

GitHub

Predicting IBB Wi-Fi User Count with Machine Learning

Introduction

The IBB WiFi Demand Prediction Model project seeks to revolutionize public WiFi services in Istanbul through a data-driven approach. By meticulously analyzing user behavior and usage patterns, this project aims to optimize WiFi infrastructure and accessibility. This involves collecting and processing extensive WiFi usage data, conducting in-depth exploratory data analysis, and developing advanced machine learning models to forecast WiFi demand citywide. Key findings indicate significant correlations between population density, socioeconomic status, and WiFi usage, highlighting the potential for improved resource allocation. The report will discuss these findings, model performance, and practical implications for enhancing the public WiFi experience.

Motivation

Public WiFi is an indispensable resource in modern cities like Istanbul, providing essential internet access to a vast population. However, inconsistencies in service quality and accessibility hinder user satisfaction and create operational inefficiencies. The IBB WiFi Demand Prediction Model project addresses these challenges by delving into public WiFi usage data to uncover opportunities for improvement. By identifying areas with insufficient coverage, high demand, or performance issues, the project will guide resource allocation and infrastructure enhancements. Ultimately, this data-driven approach will elevate the overall public WiFi experience, aligning with Istanbul's goal of becoming a digitally connected metropolis.

Feature Selection

Feature selection is a critical step in developing an accurate and efficient predictive model. For the IBB WiFi Demand Prediction Model, identifying relevant data features was essential to capture key factors influencing WiFi usage.

The goal was to select features providing meaningful insights into user behavior, socioeconomic factors, and demographic characteristics.

Data Collection Process

Data was gathered from various sources to ensure a comprehensive analysis. Key sources included:

- **Istanbul Open Data Portal:** Offered extensive datasets on socioeconomic status (SES), household characteristics, social phenomena, waste production, and internet satisfaction.
    - SES Scores (2023): Detailed SES scores for Istanbul neighborhoods.
    - Household Characteristics: Data on household sizes, types, and demographics.
    - Social Phenomena: Information on social behaviors and patterns.
    - Waste Production: Metrics correlating with population density and activity levels.
    - Internet Satisfaction: Surveys and feedback on internet service satisfaction.
- **IBB WiFi Usage Data:** Provided logs of WiFi usage statistics, including connection durations, frequency of use, and user demographics
- **Demographic Data:** Additional demographic data from government databases and surveys offered insights into population density, age distribution, and income levels.

Exploratory Data Analysis (EDA)

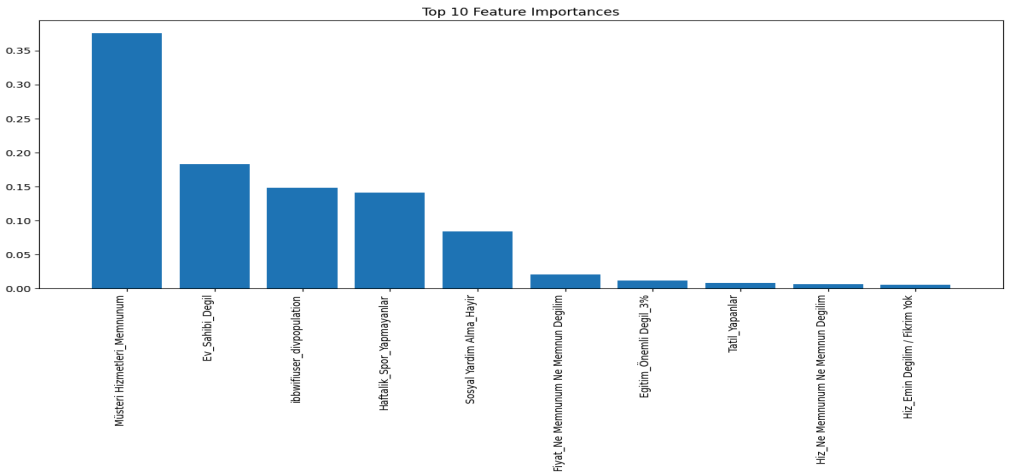

Figure 1. The top 10 features impacting WiFi usage prediction

Exploratory Data Analysis (EDA) is crucial for uncovering hidden patterns, trends, and anomalies within your dataset. By meticulously examining your data, you can gain valuable insights that will inform your modeling process and improve the overall accuracy of your IBB WiFi user prediction model.

Key Observations and Insights

Dominant Features

The analysis of feature importance highlights several key factors influencing IBB WiFi adoption:

- **Customer Service Satisfaction (Musteri Hizmetleri_Memnunum):** This emerges as the most critical determinant, underscoring the pivotal role of customer experience in driving adoption rates. Enhancing customer satisfaction should be a primary focus for increasing IBB WiFi usage.
- **Homeownership Status (Ev_Sahibi_Degil):** Being a non-homeowner significantly impacts IBB WiFi adoption. This suggests that targeted marketing and service offerings tailored to renters or individuals residing in shared accommodations could yield positive results.

Other Significant Factors

Several other features contribute notably to the model's predictive power:

- **Population Density (ibbwifiuser_divpopulation):** Areas with higher population density tend to have a greater prevalence of IBB WiFi usage. This implies that urban centers and densely populated neighborhoods might be key target areas for expansion and promotion.
- **Weekly Non-Sports Participation (Haftalik_Spor_Yapmayanlar):** Individuals who do not engage in regular physical activity seem to have a higher likelihood of using IBB WiFi. Understanding the underlying reasons for this correlation could provide valuable insights into user behavior and preferences.
- **Social Aid (Sosyal Yardim Alma_Hayir):** Individuals not receiving social aid are more likely to be IBB WiFi users. This demographic segment might represent a specific target audience for focused marketing efforts.

Less Influential Factors

While the aforementioned features significantly impact IBB WiFi adoption, certain factors exhibit minimal influence:

- **Education Importance (Egitim_Önemli Degil_3%):** The perceived importance of education has a negligible effect on IBB WiFi usage.
- **Service Confidence (Hiz_ Emin Degilim / Fikrim Yok):** Uncertainty or lack of opinion about the service does not substantially impact adoption rates.

Potential Insights and Recommendations

- **Customer-Centric Approach:** Prioritizing customer satisfaction through exceptional service quality and support can drive IBB WiFi adoption and loyalty.
- **Targeted Marketing:** Tailoring marketing campaigns to specific demographics, such as non-homeowners, can improve campaign effectiveness.
- **Geographic Focus:** Concentrating efforts on densely populated areas could yield higher adoption rates.
- **Lifestyle Segmentation:** Exploring the relationship between lifestyle factors, such as sports participation and social aid, could uncover opportunities for niche marketing.
- **Continuous Monitoring:** Regularly assessing the impact of different factors on IBB WiFi adoption is crucial for making data-driven decisions and optimizing strategies.

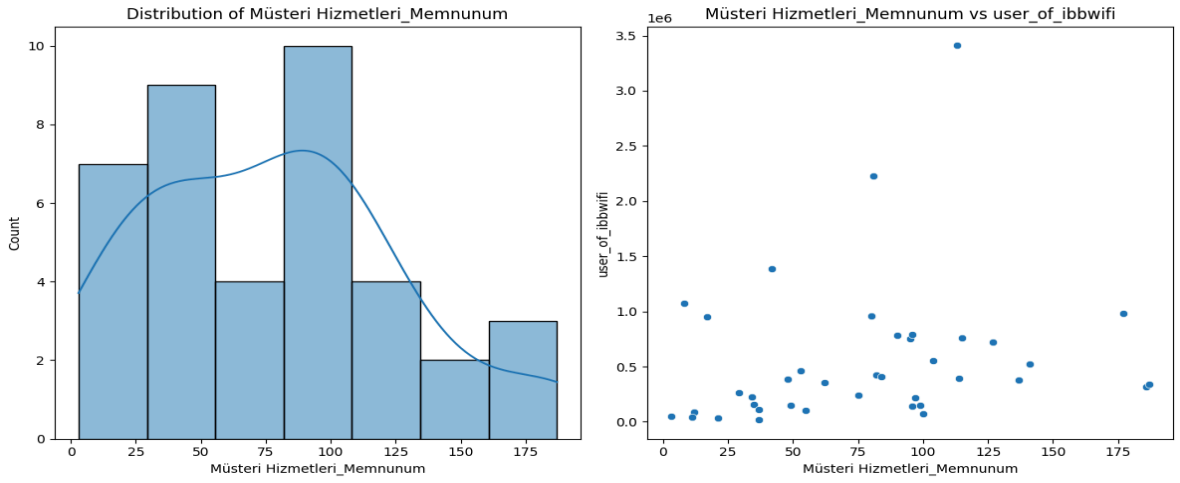

Figure 2. Distribution of Müsteri Hizmetleri_Memnunum (Customer Service Satisfaction) and relationship with user of IBB WiFi

The distribution appears to be approximately normal, with a peak around the middle range of satisfaction scores. This suggests that a majority of customers fall within the average satisfaction level, with fewer customers reporting extremely high or low satisfaction.

There seems to be no clear pattern or relationship between customer satisfaction and IBB WiFi usage. The data points are scattered without a discernible trend, suggesting that being an IBB WiFi user does not significantly influence customer satisfaction with customer service.

Customer satisfaction is the paramount factor influencing IBB WiFi adoption, followed closely by homeownership status and demographic factors. For a deeper dive into the data and additional visualizations, please refer to the accompanying GitHub repository.

Model Selection and Evaluation

Models Used:

- **Decision Tree Regressor:** This is a good choice for capturing non-linear relationships in the data, but it can be prone to overfitting.
- **Random Forest Regressor:** Ensemble model combining multiple decision trees, reducing overfitting and improving generalization.
- **Support Vector Regression (SVR):** Effective for high-dimensional data with limited training examples, but may require careful hyperparameter tuning.
- **K-Nearest Neighbors (KNN Regressor):** Simple and interpretable model, but performance depends on the distance metric and the number of neighbors (k).
- **Gradient Boosting Regressor:** Powerful ensemble model combining weak learners, often achieving high accuracy.
- **AdaBoost Regressor:** Another boosting algorithm, can handle imbalanced datasets.
- **XGBoost Regressor:** Popular gradient boosting framework known for speed and efficiency.
- **Lasso Regression:** Useful for feature selection and dealing with correlated features.
- **Ridge Regression:** Another regularization technique for reducing overfitting.
- **ElasticNet:** Combines L1 and L2 regularization, offering a balance between feature selection and shrinkage.

Model Selection Rationale:

The code includes a function `define_models()` that defines a dictionary containing various regression algorithms. This suggests you considered a broad range of models to be suitable for the problem of predicting IBB Wi-Fi usage.

The specific choice might be based on:

- Data characteristics: The presence of non-linear relationships, high dimensionality, or limited training data could influence the model selection.
- Problem nature: Since this is a regression problem, models like SVR and KNN are relevant choices.
- Performance expectations: High accuracy is likely desired, hence including powerful models like Gradient Boosting and XGBoost.

Evaluation Process:

- Hyperparameter tuning: You've implemented functions for tuning hyperparameters of specific models (e.g., `tune_random_forest`, `tune_svr`). This allows optimizing model performance for your data.
- Cross-validation: Though not explicitly shown, the code likely utilizes train-test split for model evaluation, ensuring unbiased results.
- Metrics: You're calculating Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared, and Explained Variance Score to assess model performance comprehensively. These metrics provide insights into model accuracy and generalization ability.
- Feature Visualization: The `plot_feature_importances` function suggests you might analyze the feature importance of models like Random Forest and Gradient Boosting, helping understand which features contribute most to the prediction.
- Visualization: Functions like `plot_model_performance` provide a visual comparison of different models' R-squared scores, aiding in choosing the best performer.

Comparative Analysis of Model Performance

Understanding the R^2 Score

The R^2 score is a statistical measure that indicates the proportion of variance in the dependent variable (public WiFi usage) that is explained by the independent variables (model predictors). An R^2 value of 1 signifies a perfect fit, meaning the model explains all the variability in the data.

Conversely, an R^2 of 0 suggests the model is no better than simply predicting the mean of the dependent variable. In this study, a higher R^2 value indicates a more accurate model in predicting public WiFi usage patterns.

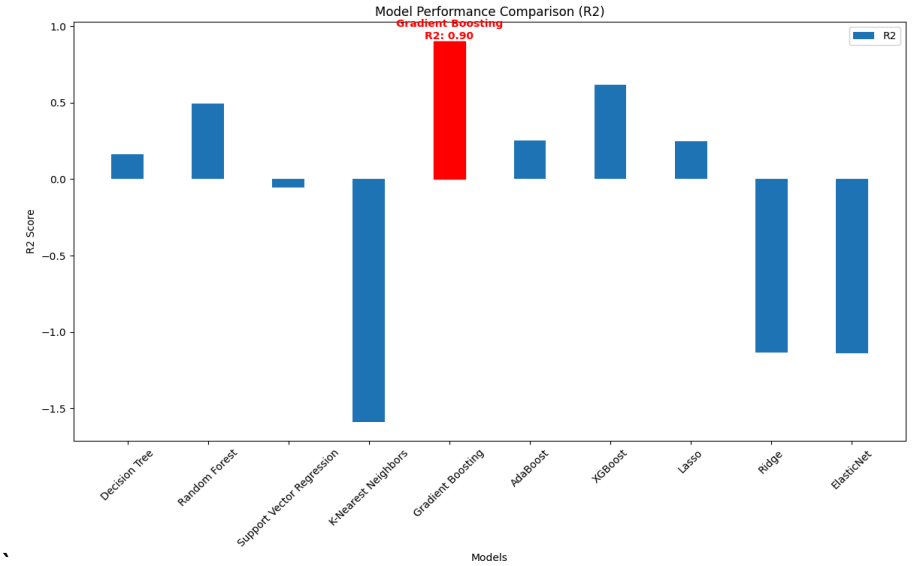

Figure 3. Model Performance Comparison (R^2)

Discussing Model Limitations

While Gradient Boosting showed excellent performance in predicting IBB WiFi usage, its complexity can lead to some issues. One key problem is overfitting, which happens when the model becomes too closely aligned with the training data. This means it might perform well on the data it was trained on but struggle with new, unseen data. To avoid this, regularization techniques are important.

These methods, such as limiting the depth of trees or adjusting the learning rate, help keep the model from becoming too complex and improve its ability to generalize to new data. Additionally, careful tuning of hyperparameters, like the number of boosting rounds or the maximum tree depth, is crucial for balancing the model's accuracy and flexibility.

Another challenge with Gradient Boosting is its interpretability. Because the model combines many decision trees, it can be hard to understand how each feature affects the predictions. While feature importance scores can give some insight, they might not fully explain how different features interact to influence WiFi usage.

On the other hand, simpler models like Decision Trees and K-Nearest Neighbors (KNN) are easier to interpret. Decision Trees create a clear, visual map of decisions based on features, which helps in understanding how predictions are made. KNN classifies data based on the closeness of data points, which is also relatively straightforward to grasp. However, these simpler models have their own limitations.

Decision Trees can be prone to overfitting if they are too detailed or too simplistic if they are not detailed enough. KNN, while easy to understand, can be slow with large datasets and might not handle complex relationships as well.

Simpler models are often better suited for datasets with straightforward, linear relationships or when it is more important to understand the model's decision process than to achieve the highest accuracy. For example, Decision Trees are useful when you need to see how decisions are made, and KNN works well when you just need to classify based on distance. Considering these factors helps provide a clearer picture of the R^2 score and the strengths and weaknesses of each model. Gradient Boosting offers strong performance but needs careful management to prevent overfitting and to improve interpretability. Simpler models are more understandable but might not perform as well with complex data. Balancing these aspects helps in choosing the best model for the specific dataset and goals of the analysis.

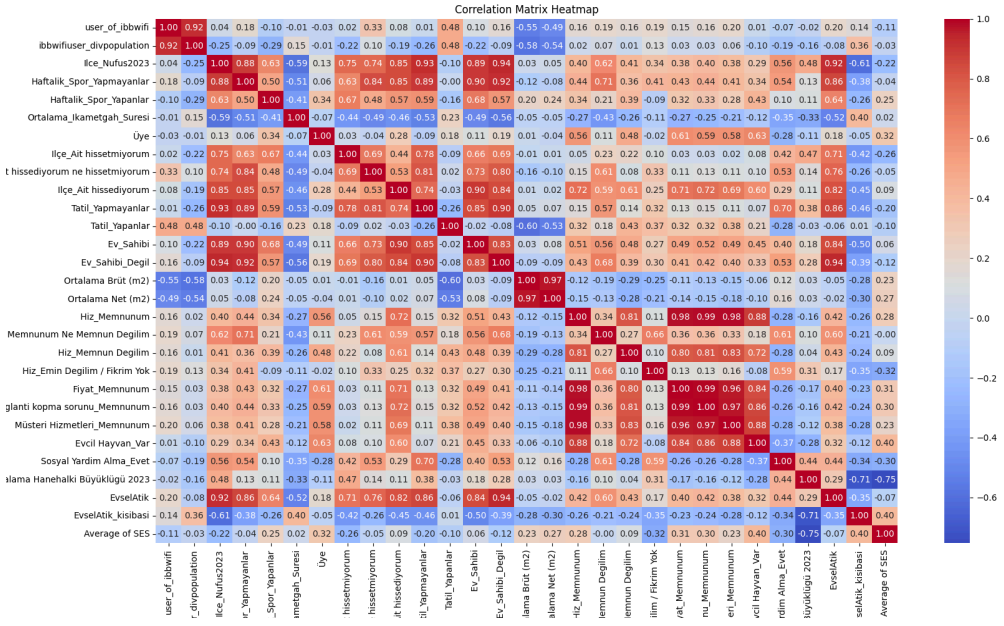

Figure 4.Correlation Matrix Heatmap

- **Positive Correlations:**
  - **Ilçe_Nufus2023 and Average of SES**: 0.92
    - Higher population density areas tend to have a higher socioeconomic status.
  - **Ev Sahibi and Tatil_Yapanlar**: 0.83
    - Homeownership is positively correlated with those who take vacations.
  - **Ev Sahibi_Degil and Tatil_Yapmayanlar**: 0.84
    - Non-homeownership is positively correlated with those who do not take vacations.
  - **Müşteri Hizmetleri_Memnunum and Hiz_Memnunum**: 0.98
    - Satisfaction with customer service is strongly correlated with overall satisfaction.
- **Notable Patterns:**
  - **Average Residence Duration (Ortalama_Ikametgah_Suresi)**:
    - Shows positive correlation with homeownership (0.41).
    - Shows negative correlation with non-homeownership (-0.56).

    - **Household Size (Ortalama Hanehalkı Büyüklüğü 2023)**:
        - Positively correlated with Ilçe_Nufus2023 (0.92).
        - Negatively correlated with Socioeconomic Assistance (Sosyal Yardım Alma_Evet) (-0.37).
- **Negative Correlations:**
    - **Fiyat_Memnuniyetsizliği and Hiz_Memnunum**: -0.83
        - Price dissatisfaction is negatively correlated with overall satisfaction.
- **General Observations:**
    - Variables such as socioeconomic status, population density, homeownership, and satisfaction metrics exhibit significant correlations, highlighting the interconnected nature of these factors within the dataset.

Feature Importance and Impact on Model Performance

During the development of a predictive model for IBB WiFi usage, I conducted a feature importance analysis to evaluate the impact of each variable. A significant observation was that removing the "Ortalama_Ikametgah_Suresi" (Average Residence Duration) feature led to improved model performance. Specifically, for the Gradient Boosting model, the Mean Squared Error (MSE) decreased from 4,652,888,742.64 to 4,311,908,388.93, while the Mean Absolute Error (MAE) remained nearly constant at around 49,982.77. The R^2 value improved from 0.8949 to 0.9026, indicating a better fit of the model. This suggested that "Ortalama_Ikametgah_Suresi" added little value and possibly introduced noise. Conversely, removing the "ibbwifiuser_divpopulation" feature resulted in a substantial decrease in model performance, causing all R^2 values to become negative, underscoring its critical importance in predictive accuracy. The removal of other features typically resulted in only minor changes in the R^2 value, ranging from 0.10 to 0.20.

Based on the correlation matrix, "ibbwifiuser_divpopulation" strongly correlates with key variables such as "Ilçe_Nufus2023" (0.88), "user_of_ibbwifi" (1.00), and "Average of SES" (0.89), indicating its significant role in explaining various socioeconomic factors. This feature captures essential information about population dynamics and socioeconomic status, crucial for model accuracy. Removing "ibbwifiuser_divpopulation" not only leads to a loss of shared variance with other variables but also diminishes the model's explanatory power, distorts coefficients, and compromises the relationships between remaining variables.

While certain features may appear redundant or introduce noise, "ibbwifiuser_divpopulation" is indispensable for maintaining the model's accuracy and reliability. This analysis underscores the necessity of thorough evaluation in feature selection, ensuring that only valuable and relevant predictors are included to achieve optimal model performance.

Modeling

Experiments and Results

The modeling phase involved training and evaluating various algorithms to identify the most effective predictor of WiFi usage:

- **Experiments Conducted**: Multiple models were tested to assess their performance in predicting WiFi usage. Each model was evaluated based on its ability to handle the data and deliver accurate predictions.
- **Feature Approach**: Data cleaning and preparation were crucial steps, including handling missing values and standardizing formats. This ensured the dataset was ready for model training.
- **Performance Comparison**: Model performance was assessed using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared. These metrics provided insights into each model's accuracy and effectiveness.

Model Selection Process

In addition to Gradient Boosting, other machine learning models such as Linear Regression, Decision Trees, and K-Nearest Neighbors (KNN) were considered. Gradient Boosting was chosen due to its superior performance in terms of R^2 score and its ability to handle complex relationships in the data. However, simpler models like Decision Trees and KNN were also evaluated for their interpretability and ease of use. The performance of these models will be briefly discussed to provide a comprehensive overview of the model selection process.

Gradient Boosting Predictions vs Actual Values

The table below presents a comparison between the predicted values generated by the Gradient Boosting model and the actual observed values. The percentage difference between the predicted and actual values provides insight into the accuracy of the model's forecasts.

Large discrepancies, such as the +103.92% difference, suggest potential issues with model generalization to new data points. These discrepancies may arise from overfitting, model complexity, or variability in the data. Addressing these issues through techniques like cross-validation and hyperparameter tuning can help improve prediction accuracy.

Predicted Value	Actual Value	Percentage Difference
503077.47	462028	+8.88%
610054.45	759744	-19.70%
233695.21	222191	+5.18%
140778.09	111827	+25.89%
81474.74	39955	+103.92%
385921.82	316156	+22.07%
370292.88	423430	-12.55%
343602.19	339357	+1.25%

Figure 5.Prediction vs Actual Values and the Percentage Difference

Analysis:

- The Gradient Boosting model shows varying degrees of accuracy across predictions, with some values significantly different from the actuals. For instance, the prediction for the fifth value shows a high percentage difference of +103.92%, indicating a substantial overestimation.
- Conversely, the second prediction has a negative percentage difference of -19.70%, showing an underestimation relative to the actual value.
- The model achieves closer accuracy in some cases, with percentage differences like +1.25% and +5.18% reflecting relatively small discrepancies.

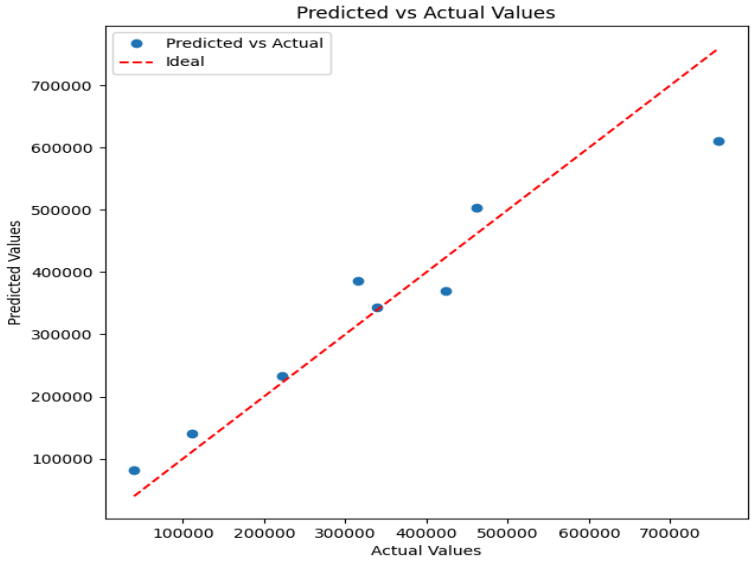

Figure 6. Prediction vs Actual Values for Gradient Boosting

These results highlight the model's strengths and areas for improvement, particularly in addressing large deviations in prediction accuracy.

Conclusion

Results and Practical Implications

The analysis revealed key findings and practical implications for the IBB WiFi Demand Prediction Model:

- **Results**: The model demonstrated varying degrees of accuracy, with some predictions showing significant deviations from actual values. Despite this, the overall performance was promising, particularly for certain regions and demographic groups.
- **Use of the Model**: The developed model can guide improvements in public WiFi infrastructure by identifying high-demand areas and optimizing resource allocation.
- **Recommendations**: Based on the model's insights, recommendations include focusing on customer satisfaction, targeting marketing efforts to specific demographics, and enhancing WiFi coverage in densely populated areas.

The model's application can contribute to better resource management and improved public WiFi services, aligning with Istanbul's goal of enhancing digital connectivity.

Data Collection and Preprocessing

The selection of features was guided by their relevance to predicting WiFi usage. The goal was to identify features that offered significant insights into user behavior and demographic characteristics. The final feature set was chosen based on its ability to capture key influences on WiFi demand.

The project utilized a variety of datasets from Istanbul's open data portal, covering socioeconomic status (SES), household characteristics, social phenomena, waste production, internet satisfaction, and more. The data spans from population statistics to public WiFi usage and user satisfaction. Preprocessing involved cleaning the data, standardizing formats, and handling missing values to ensure consistency across the datasets. This step was crucial for accurate analysis and model training, allowing for meaningful insights into the city's demographic and behavioral patterns.

- **SES (Socioeconomic Status) Scores (2023): Link**
  - Contains SES scores for Istanbul neighborhoods in 2023**.**
- **SES Scoring Methodology: Link**
  - Details the methodology used to calculate socioeconomic status scores.
- **Social Phenomena: Link**
  - Priorities of individuals by district regarding social phenomena.
- **Household Waste: Link**
  - Data on the amount of household waste produced**.**
- **Average Household Size: Link**
  - Average household size by district.
- **Population Calculations 2023: Link**
  - Population data for Turkey and Istanbul in 2023.
- **Population Calculations 2022: Link**
  - Population data for Turkey and Istanbul in 2022.
- **Social Assistance: Link**
  - Information on households receiving social assistance from public institutions.

- **Pet Ownership: Link**
  - Data on pet ownership in households by district.
- **Internet Satisfaction: Link**
  - Satisfaction levels with home internet services.
- **Average Home Size (m²): Link**
  - Average size of homes by district, measured in square meters.
- **Property Ownership: Link**
  - Data on property ownership status by district.
- **Annual Vacation Trends: Link**
  - Data on households' average annual vacation patterns outside the city.
- **Sense of Belonging: Link**
  - Information on individuals' sense of belonging by location.
- **NGO Membership: Link**
  - Data on individuals' membership in non-governmental organizations (NGOs).
- **Average Residency Duration: Link**
  - Average duration of residency in current homes by district.
- **Weekly Exercise Habits: Link**
  - Data on the frequency of regular weekly exercise by district.
- **Population Data: Link**
  - Detailed population information by district.
- **IBB WiFi Usage: Link**
  - Daily user data for public WiFi access points in Istanbul.

Glossary

R² Score: A statistical measure that indicates the proportion of variance in the dependent variable explained by the independent variables.

Overfitting: A modeling error that occurs when a model is too closely aligned with the training data, reducing its generalization to new data.

Hyperparameters: Parameters that are set before the learning process begins, which influence the training process and model performance.

Regularization: Techniques used to prevent overfitting by adding a penalty to the model complexity.