# FATER GEO-ANALYTICS CHALLENGE

by

Raza Mehar, Pujan Thapa, Syed Najam Mehdi

**A report on analyzing Fater’s diaper market potential and enhancing revenue estimation for Naples stores in the entire province: A Socio-Demographic, Territorial, and Points of Interest Perspective.**

Course: Statistical Learning & Data Analysis

Department of Engineering

University of Naples Federico II

February 2024

## Executive Summary

The FATER Geo-Analytics Challenge explores the estimation of potential revenue for Naples stores throughout the entire province, with a focus on the Italian diaper market (in 1000s). The study incorporates a detailed approach, considering socio-demographic factors, geographical features, and points of interest, to extract insights for strategic decision-making in the competitive retail landscape.

A primary revelation from our analysis is the influence of hypermarkets and larger stores in shaping potential revenue dynamics. These establishments demonstrate a considerable capacity for customer retention and sales generation. Their expansive footprint in the market positions them as pivotal players, warranting strategic marketing emphasis to leverage their inherent potential fully.

A critical determinant identified in our study is the role of parking availability, showcasing an impact on potential values. Stores equipped with parking facilities exhibit higher potential value, underscoring the significance of convenient access for customers. This insight enables businesses to tailor promotional strategies, highlighting the presence of parking facilities as a compelling feature for attracting and retaining customers.

Temporal dynamics form a key facet of our findings, with distinct gravitational patterns emerging during weekends and specific evening time slots, particularly from 17:00 to 20:00. This temporal analysis provides businesses with a strategic edge, allowing them to optimize promotional campaigns during peak hours, aligning with heightened customer gravitation.

Delving into demographics, our study identifies specific age groups displaying a higher inclination towards stores, notably individuals aged 41-50 and those over 60. Additionally, a gender-based trend surfaces, with males exhibiting a higher average gravitation towards stores.

In summary, the FATER Geo-Analytics Challenge delivers actionable recommendations for strategic marketing focus, targeted promotions, and an emphasis on parking facilities. This comprehensive roadmap equips businesses to navigate the landscape of the Italian diaper market, unlocking the full potential for revenue maximization.

Table of Contents

[Methodology 5](#_Toc163159826)

[The Datasets 5](#_Toc163159827)

[Spatial Exploration 6](#_Toc163159828)

[Data Preparation 7](#_Toc163159829)

[Exploratory Data Analysis (EDA) 9](#_Toc163159830)

[Miscellaneous Analysis 9](#_Toc163159831)

[Model Selection and Evaluation 10](#_Toc163159832)

[Feature Importance Analysis 10](#_Toc163159833)

[Results and Exploration 11](#_Toc163159834)

[Descriptive Statistics 11](#_Toc163159835)

[Univariate Analysis 11](#_Toc163159836)

[Bivariate Analysis 12](#_Toc163159837)

[Hypothesis Testing 13](#_Toc163159838)

[Correlation Analyses 14](#_Toc163159839)

[Population Gravitation Analysis 14](#_Toc163159840)

[Contingency Table 15](#_Toc163159841)

[Geo-Spatial Analysis 16](#_Toc163159842)

[Pareto Analyses 17](#_Toc163159843)

[Model Selection and Hyperparameter Tuning with Cross Validation 18](#_Toc163159844)

[Performance Evaluation of the Model 19](#_Toc163159845)

[Feature Importance Analysis 20](#_Toc163159846)

[Conclusion 21](#_Toc163159847)

[Findings and Observations: 22](#_Toc163159848)

[Recommendations 23](#_Toc163159849)

[Complimentary 24](#_Toc163159850)

[Population Dynamics up till 2043 24](#_Toc163159851)

[Population and Household Projections for South Italy 24](#_Toc163159852)

[Strategies 25](#_Toc163159853)

[Bibliography 26](#_Toc163159854)

[Libraries & Modules 27](#_Toc163159855)

# Methodology

The methodology adopted a systematic framework, exploring the datasets, encompassing spatial exploration, data preparation, exploratory data analysis (EDA), and model selection and evaluation.

### The Datasets

The FATER datasets comprise essential information on various aspects, including stores (1270 rows, 11 columns), the geometry of microcode polygons (10077 rows, 2 columns), demographics (10077 rows, 14 columns), and population gravitation (806160 rows, 6 columns). These datasets as shown in Table 1.0, collectively form the foundation for our comprehensive analysis, offering insights into store characteristics, spatial details, demographic trends, and population gravitation patterns.

|  |  |  |  |
| --- | --- | --- | --- |
| Stores Dataset | Shapes Dataset | Demographics Dataset | Gravitation Dataset |
| store\_ID: int  store\_name: chr  store\_type: chr  store\_size: num  parking: chr  address: chr  lat: num  long: num  comune: chr  province: logi  potential: num | geometry: chr  microcode: num | microcode: num  district: chr  province: chr  region: chr  population: int  population\_m: int  population\_f: int  population\_age\_00\_04\_yr: int  population\_age\_05\_14\_yr: int  population\_age\_15\_34\_yr: int  population\_age\_35\_44\_yr: int  population\_age\_45\_54\_yr: int  population\_age\_55\_64\_yr: int  population\_age\_65\_up\_yr: int | Unamed: int  microcode: num  daytype: int  time\_slot: int  datatype: chr  media\_annuale: num |

*Table 1.0: The four raw datasets used for the analysis.*

Additional datasets include the Population by sex and age class – Municipalities Campania (4160 row, 8 columns) and Population change components – Municipalities Campania (91728 row, 9 columns) from The National Institute of Statistics. These datasets as shown in Table 1.1, help us understand the population dynamics of Naples from present up till 2043.

|  |  |
| --- | --- |
| Population by sex and age class | Population change components |
| Year: int  Municipality Code: int  Municipality: object  Province code: int  Province: object  Population at start: int  Live births: int  Population at end: int | Year: int  Municipality Code: int  Municipality: object  Codice provincia: int  Age: object  Males: int  Females: int  Total: int |

*Table 1.1: The four raw datasets used for the analysis.*

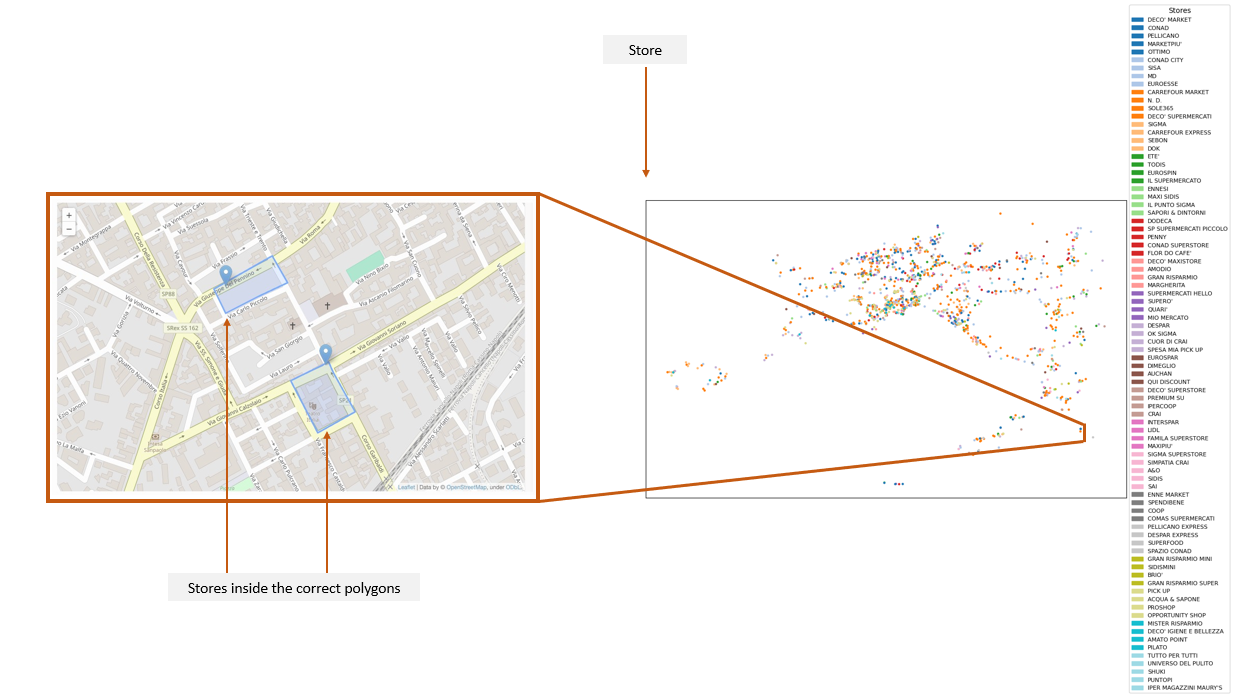
### Spatial Exploration

Initiating a meticulous spatial exploration, stores were strategically georeferenced within polygons to extract precise micro-codes. Leveraging the Shapely library, this step ensured the requisite spatial accuracy for subsequent analyses. Initially, a cartographic overlay was performed, plotting both stores and polygons on maps (refer to Figure 1.0) to verify spatial relationships. Once confirmed that polygons encapsulate distinct stores, subsequent spatial processing involves the precise mapping of relevant stores within the designated polygons (as illustrated in Figure 1.1). Following the spatial processing and plotting, all datasets were integrated through a spatial join using the unique microcode ID for a comprehensive analysis.

A screenshot of a map

Description automatically generated

*Figure 1.0: Plotting the stores and polygons on the maps to explore whether there is a mapping between them.*



*Figure 1.1: Plotting the stores inside the correct polygons to extract accurate micro-codes.*

### Data Preparation

The pre-processing phase was crucial for data integrity. It involved cleaning and merging datasets and translating column names to English for standardization. To handle missing values, the median imputation method was used, with a specific focus on the potentially important 'annual average' feature.

Table 1.1 offers a detailed description of each categorical feature and its values.

|  |  |
| --- | --- |
| Categorical Feature | Description |
| store\_ID | Unique store identification number |
| store\_type | Type of the store: (IPR: Hypermarket; SUP: Supermarket; LIS; Libero Servizio; SSD: Drugstore; DIS: Discount Store) |
| parking | Yes, if parking is available, else No |
| comune | Municipality where the store is situated |
| microcode | ISTAT microcell |
| daytype | 1: weekday; 2: weekend |
| time\_slot | 2: 07 am – 10 am; 3: 10 am – 13 pm; 4: 13 pm – 14 pm; 5: 14 pm – 17 pm; 6: 17 pm – 20 pm |
| datatype | F1: population under 18 years; F2: population aged 18 – 30 years; F3: population aged 31 – 40 years; F4: population aged 41 -50 years; F5: population aged 51 – 60 years; F6: population over 60 years; Gf: female population; Gm: male population |

*Table 1.2: Analyzing categorical features scale.*

Table 1.2 offers a detailed description of each numerical feature.

|  |  |
| --- | --- |
| Numerical Feature | Description |
| store\_size | Size of the store |
| potential | Italian diaper market (In 1000s) |
| population | Total population |
| population\_m | Male population |
| population\_f | Female population |
| population\_age\_00\_04\_yr | - |
| population\_age\_05\_14\_yr: | - |
| population\_age\_15\_34\_yr | - |
| population\_age\_35\_44\_yr | - |
| population\_age\_45\_54\_yr | - |
| population\_age\_55\_64\_yr | - |
| population\_age\_65\_up\_yr | - |
| annual average | the yearly average of people gravitating by the combination of microcode, daytype, time slot, and datatype |

*Table 1.3: Analyzing numerical features Scale.*

Tables 1.2 and 1.3 show the count of missing values identified in categorical and numerical features.

|  |  |
| --- | --- |
| Categorical Feature | Count of missing values |
| microcode | 14 |
| daytype | 14 |
| time slot | 14 |
| datatype | 14 |

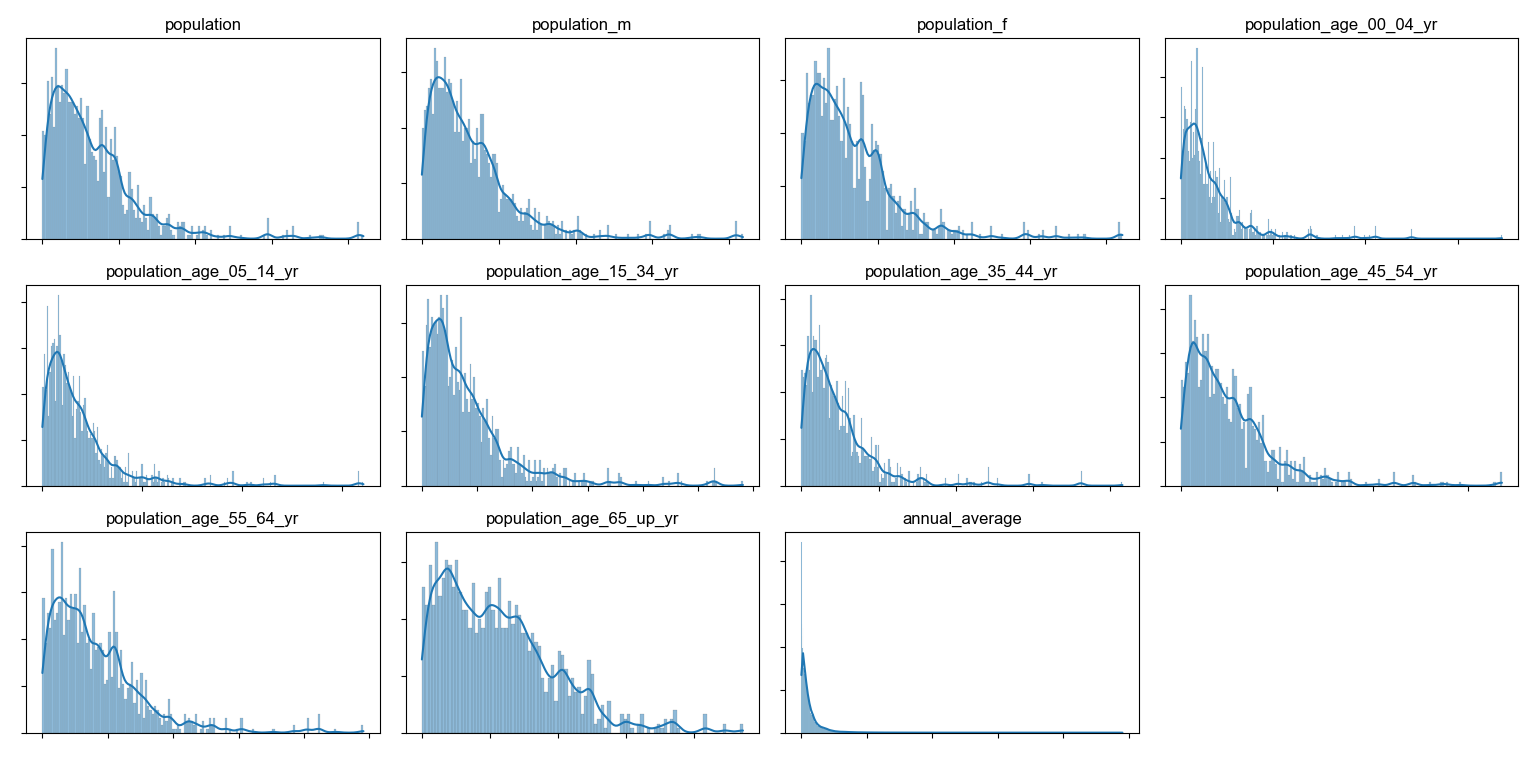
*Table 1.4: Count of missing values in Categorical Features.*

|  |  |
| --- | --- |
| Numerical Feature | Count of missing values |
| population | 14 |
| population\_m | 14 |
| population\_f | 14 |
| population\_age\_00\_04\_yr | 14 |
| population\_age\_05\_14\_yr | 14 |
| population\_age\_15\_34\_yr | 14 |
| population\_age\_35\_44\_yr | 14 |
| population\_age\_45\_54\_yr | 14 |
| population\_age\_55\_64\_yr | 14 |
| population\_age\_65\_up\_yr | 14 |
| annual average | 2145 |

*Table 1.5: Count of missing values in Numerical Features.*

The missing values account for **2.14%** of total rows for the *annual average* and **0.1%** for other features. Despite **100,494** rows in the dataset, the missing values are relatively insignificant. However, given the potential importance of the *annual average* feature, a decision has been made to impute the missing values in the numerical features.

The following figure depicts histograms used to assess normality in the numerical features with missing values.



*Figure 1.2: Plotting the Histograms of the numerical features with missing values to check for normality.*

Upon observing histograms in Figure 1.2, it has been determined that the features do not adhere to a normal distribution. As a result, missing values will be imputed using the median. The missing categorical values are imputed using the mode.

There are no duplicate values in the data.

### Exploratory Data Analysis (EDA)

Descriptive statistics provided an initial overview of numerical features. Employing univariate and bivariate analyses, both categorical and numerical features were visually explored, revealing intricate relationships and potential predictors.

Hypothesis testing played a crucial role in assessing the statistical significance of differences in potential values among store types and parking statuses. Since the data is not normally distributed, hence non-parametric tests such as Kruskal-Wallis and Mann-Whitney U tests. The Kruskal-Wallis test was applied to assess whether there are significant differences in potential values concerning the five store types, and the Mann-Whitney U test was used for potential and parking status.

Linear correlation analysis was subsequently conducted to explore relationships between potential and the numerical predictors and check if there is any multicollinearity amongst features.

### Miscellaneous Analysis

A contingency table analysis was conducted to explore the association of potential bins: Low, Medium, and High with store type, store size, and parking status, aiming to derive further actionable insights. Concurrently, an investigation into population gravitation was carried out, delving into the dynamics concerning day type, time slot, and demographics to glean valuable insights. Concurrently, The geospatial analysis of stores in correlation with their potential unraveled insightful observations. Applying the Pareto analysis, in alignment with the 80-20 principle, we identified stores and store types that wield the most significant potential.

### Model Selection and Evaluation

To enhance the predictive capabilities of the analysis, ensemble models were employed incorporating decision tree, boosting, and bagging. Decision Tree was used for their interpretability and computational efficiency. Bagging was used to reduce variance (overfitting) while boosting was employed to reduce bias (underfitting). Furthermore, ensembles are less sensitive to outliers. Specifically, Random Forest, Gradient Boosting, and Adaptive Boosting regressors using the Scikit-Learn library in Python were utilized.

Hyperparameter tuning was performed using a grid search approach with k-fold cross-validation to find the optimal set of hyperparameters for each model.

Afterward, based on the root mean squared error (RMSE) and standard deviation derived from cross-validation, the best-performing model based on the identified optimal set of hyperparameters was selected for further analysis.

Mean Squared Error (MSE), Mean Absolute Error (MAE), and coefficient of determination (R2) were employed to measure the model’s performance. The Mean Squared Error (MSE) is employed to rigorously penalize significant errors, whereas the Root Mean Squared Error (RMSE) serves the dual purpose of penalizing large errors while preserving consistency in units with the target variable. Conversely, the Mean Absolute Error (MAE) is utilized to assess model performance, accounting for the inherent noise present in the data.

### Feature Importance Analysis

The feature importance provided by the Random Forest model was explored. This analysis offered insights into which features are most influential in making predictions.

## Results and Exploration

### Descriptive Statistics

Descriptive statistics were conducted on numerical features, revealing variations in the mean and standard deviation of features.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Numerical Feature | Min | Q1 | Median | Mean | Q3 | Max | SD |
| store size | 100.0 | 150.0 | 250.0 | 413.6 | 480.0 | 11978.0 | 641.8 |
| population | 0.0 | 253.0 | 513.5 | 655.2 | 891.0 | 4197.0 | 582.7 |
| population\_m | 0.0 | 122.0 | 249.0 | 318.4 | 425.0 | 2089.0 | 289.5 |
| population\_f | 0.0 | 133.0 | 270.5 | 336.8 | 465.0 | 2108.0 | 294.2 |
| population\_age\_00\_04\_yr | 0.0 | 11.0 | 22.0 | 30.5 | 39.0 | 348.0 | 33.1 |
| population\_age\_05\_14\_yr: | 0.0 | 26.0 | 52.0 | 72.5 | 91.0 | 642.0 | 77.3 |
| population\_age\_15\_34\_yr | 0.0 | 59.0 | 122.0 | 164.4 | 214.0 | 1163.0 | 161.2 |
| population\_age\_35\_44\_yr | 0.0 | 35.0 | 70.0 | 93.8 | 122.0 | 831.0 | 93.3 |
| population\_age\_45\_54\_yr | 0.0 | 36.0 | 78.0 | 100.8 | 140.0 | 670.0 | 91.3 |
| population\_age\_55\_64\_yr | 0.0 | 31.0 | 64.0 | 81.5 | 112.0 | 491.0 | 69.9 |
| population\_age\_65\_up\_yr | 0.0 | 45.0 | 98.5 | 111.7 | 159.0 | 472.0 | 82.5 |
| annual average | 1.0 | 25.0 | 68.0 | 122.9 | 147.0 | 4903.0 | 182.6 |
| Potential | 0.001 | 0.002 | 0.005 | 0.01 | 0.013 | 0.83 | 0.04 |

*Table 1.6: Exploring descriptive statistics pre-standardization.*

As can be seen in Table 1.5, there is a difference between the mean and standard deviation of the features.

### Univariate Analysis

The univariate analysis visualized the distribution of categorical and numerical features using histograms.

A group of blue bars

Description automatically generated with medium confidence

*Figure 1.3: Plotting the bar charts of the categorical features.*

In Figure 1.3, it is evident Libero Servizio and Supermarkets dominate in the store type category, while most stores have no parking.

A screenshot of a graph

Description automatically generated

*Figure 1.4: Plotting the histograms of the numerical features.*

Figure 1.4 establishes that none of the numerical features follow a normal distribution; specifically, store size, annual average, and potential all appear constant.

### Bivariate Analysis

Bivariate analysis explored separate relationships of numerical and categorical predictors with potential, revealing interesting insights.

It is evident from Figure 1.6 that the potential values for Hypermarkets are generally much larger and have greater variability as compared to other stores. Additionally, stores with parking exhibit somewhat higher potential values than those without parking.

An additional statistical test will be conducted to determine the statistical significance of the observed differences in potential values between store types and parking status

A screenshot of a graph

Description automatically generated

*Figure 1.5: Exploring relationships between numerical predictors and potential.*

A group of graphs and diagrams

Description automatically generated

*Figure 1.6: Exploring relationships between categorical predictors and potential.*

### Hypothesis Testing

Hypothesis testing using Kruskal-Wallis and Mann-Whitney U tests assessed the statistical significance of differences in potential values among store types and parking statuses respectively. Based on the very small p-values approaching 0 we can suggest that the differences in potential values between store type and parking are statistically significant, with more potential values lying within Hypermarkets and stores with parking. This implies that store type and parking status significantly influence potential values. The correlation between them will determine if there is any linear relationship.

***Results:***

* Kruskal-Wallis-Test: p-value: 0.0
* Mann-Whitney U Test: p-value: 0.0

### Correlation Analyses

The linear correlation analyses examined relationships between potential and various predictors using the Pearson method.

A table with numbers and letters

Description automatically generated

*Figure 1.7: Examining relationships between numerical predictors and potential.*

In Figure 1.7, it is evident that all the population-related features exhibit high correlation, suggesting the presence of multicollinearity. For subsequent analysis, only the features 'population\_m' and 'population\_f' will be considered. Additionally, a notable positive correlation is observed between the potential and store size.

### Population Gravitation Analysis

Analyses of population gravitation concerning day type, time slot, and demographics offered valuable insights for customer acquisition strategies.

A graph of different sizes of bars

Description automatically generated with medium confidence

*Figure 1.8: Average population gravitation analyses concerning day type, time slot, and demographics.*

Figure 1.9 illustrates that the gravitational pull during the 5-day workweek is like that of the 2-day weekend, implying that most people gravitate during those 2 weekend days. The time slot when the gravitational force is the strongest is from 17pm – 20pm. The demographic that gravitates the most consists of age groups 41–50 and those over 61. Males tend to gravitate towards stores the most.

### Contingency Table

Simultaneously, the contingency table of potential bins of Low, Medium, and High concerning store type, store size, and parking status provided recommendations for retaining existing customers.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Potential\  Store Type | Discount Store | Hypermarket | Libero Servizio | Drug Store | Supermarket | Total |
| Low (0.001 - 0.241) | 8242 | 160 | 53126 | 11601 | 26884 | 100013 |
| Medium (0.241 - 0.601) | 0 | 321 | 0 | 0 | 0 | 321 |
| High (0.601 - 0.841) | 0 | 160 | 0 | 0 | 0 | 160 |
| Total | 8242 | 641 | 53126 | 11601 | 26884 | 100494 |

*Table 1.7: Contingency Table of Potential concerning Store Type.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Potential\  Store Size | Small  (100 - 1000) | Medium  (1000 - 5000) | Large  (5000 - 12701) | Total |
| Low (0.001 - 0.241) | 94653 | 5360 | 0 | 100013 |
| Medium (0.241 - 0.601) | 0 | 81 | 240 | 321 |
| High (0.601 - 0.841) | 0 | 0 | 160 | 160 |
| Total | 94653 | 5441 | 400 | 100494 |

*Table 1.8: Contingency Table of Potential concerning Store Size.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Potential\  Parking | No | Yes | Total | Potential\Store Type |
| Low (0.001 - 0.241) | 62247 | 37766 | 100013 | Low (0.001 - 0.241) |
| Medium (0.241 - 0.601) | 0 | 321 | 321 | Medium (0.241 - 0.601) |
| High (0.601 - 0.841) | 0 | 160 | 160 | High (0.601 - 0.841) |

*Table 1.9: Contingency Table of Potential concerning Parking.*

Tables 1.7 – 1.9 indicate that Hypermarkets have the highest potential compared to other types of stores. Larger stores, in general, exhibit more potential. Additionally, stores with parking facilities also show increased potential.

### Geo-Spatial Analysis

Geospatial analysis of stores concerning their potential yields some insightful observations.

A map of the world

Description automatically generated

*Figure 1.9: Visualizing potential across different stores.*

|  |  |  |  |
| --- | --- | --- | --- |
| Store ID | Store Name | Comune | Store Type |
| 7161 | AUCHAN | GIUGLIANO IN CAMPANIA | IPR |
| 7816 | AUCHAN | NOLA | IPR |
| 11626 | AUCHAN | MUGNANO DI NAPOLI | IPR |
| 11412 | AUCHAN | POMPEI | IPR |
| 23157 | SPAZIO CONAD | VOLLA | IPR |
| 7815 | IPERCOOP | QUARTO | IPR |
| 10356 | IPERCOOP | AFRAGOLA | IPR |
| 20804 | SP SUPERMERCATI PICCOLO | CASTELLO DI CISTERNA | SUP |
| 7456 | DECO' SUPERSTORE | SANT'ANASTASIA | SUP |
| 4381 | CONAD SUPERSTORE | POMPEI | SUP |

*Table 2.0: Top 10 stores with the most potential.*

In the top 10 stores with the most potential, 7 are Hypermarkets, further substantiating the positive relationship identified previously in the correlation matrix between potential and Hypermarkets.

### Pareto Analyses

The Pareto analysis, following the 80-20 principle, has identified pivotal stores and store types that contribute significantly to the maximum potential. This analysis serves as a valuable supplement to the previously established findings.

A graph of a graph

Description automatically generated with medium confidence

*Figure 2.0: Visualizing stores and types that generate the most potential.*

Figure 2.2 illustrates the Pareto Charts, revealing a significant insight. Approximately 80% of the cumulative potential is derived from just 20% of the distinct stores, specifically totaling 373 out of the overall count of 1270 stores. Further details, including the list of these influential 373 stores, are provided in the detailed presentation.

Furthermore, Hyperstores, which constitute 20% of the store types, contribute to more than 80% of the potential.

### Model Selection and Hyperparameter Tuning with Cross Validation

The hyperparameter tuning process involved exploring various configurations for the Decision Tree, Random Forest, Gradient Boosting, and AdaBoost models using 80% of the data for training while reserving 20% as a hold-out set—data that the model has not seen.

These model parameters, as can be seen in Table 2.0, were subjected to k-fold cross-validation (k = 10) to identify the best-performing configurations. The choice of k = 10 aligns with established best practices, supported by empirical evidence. Ron Kohavi's experiments on diverse real-world datasets indicate that a 10-fold cross-validation strategy strikes an optimal balance between bias and variance in model assessment[[1]](#footnote-1).

|  |  |  |  |
| --- | --- | --- | --- |
| Model | n\_estimators | max\_depth | learning\_rate |
| Decision Tree | Not Applicable | None, 5 | Not Applicable |
| Random Forest | 50, 100 | None, 5 | Not Applicable |
| Gradient Boosting | 50, 100 | 3, 5 | 0.01, 0.1 |
| Adaptive Boosting | 50, 100 | Not Applicable | 0.01, 0.1 |

*Table 2.1: Configurations for the models used for hyperparameter tuning.*

After an exhaustive grid search based on the negative mean square error scoring metric, the best hyperparameters for each ensemble regressor were identified, providing a foundation for subsequent model evaluation.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | n\_estimators | max\_depth | learning\_rate |
| Decision Tree | Not Applicable | None | Not Applicable |
| Random Forest | 100 | None | Not Applicable |
| Gradient Boosting | 100 | 5 | 0.1 |
| Adaptive Boosting | 100 | Not Applicable | 0.01 |

*Table 2.2: Optimal parameters identified through Grid Search.*

Following the hyperparameter tuning, the models underwent rigorous evaluation using a 10-fold cross-validation approach. The root mean squared error (RMSE) was employed as a key metric to assess the predictive performance of each model. The RMSE values indicate how well each model is performing in terms of predicting the target variable. The lower the RMSE, the better the model's predictions align with the actual values. Additionally, the standard deviation of RMSE provides a measure of variability in the model's performance across different folds. A lower standard deviation suggests more consistent performance.

|  |  |  |
| --- | --- | --- |
| Model | Root Mean Square Error | Standard Deviation |
| Decision Tree | 0.0009674833109087932 | 0.0005208617581119908 |
| Random Forest | 0.0008442124556278822 | 0.0005500335117800331 |
| Gradient Boosting | 0.005899866091907829 | 0.0002649940076146541 |
| Adaptive Boosting | 0.014027114258543435 | 0.0003043495517493041 |

*Table 2.3: RMSE and standard deviation of each model after 10-fold cross-validation.*

As depicted in Table 2.2, Random Forest exhibits the lowest compared to the other two models, suggesting superior performance in our evaluation.

### Performance Evaluation of the Model

The optimal model was then trained on the entire 100% of the training data. Subsequently, the model underwent evaluation on the reserved holdout set to assess its performance. Mean Squared Error (MSE), Mean Absolute Error (MAE), and coefficient of determination (R2) metrics were employed to evaluate the Random Forest Regressor's performance.

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Training Data | Holdout Data |
| Mean Squared Error | 0.0000002354 | 0.0000004056 |
| Mean Absolute Error | 0.0000318994 | 0.0000565263 |
| Coefficient of Determination | 0.9998815325 | 0.9997456078 |

*Table 2.4: Model Evaluation using Mean Squared Error, Mean Absolute Error, and Coefficient of Determination.*

The Mean Squared Error (MSE) and Mean Absolute Error (MAE) values for both sets are quite low, suggesting that the model is making accurate predictions with both large and small errors respectively.

The R2 values for both the training and holdout sets are very high (close to 1), indicating that the model is explaining a significant amount of variance in both datasets which is the goodness of fit.

The standard error of the predicted values is 0.00028, which indicates that predicted values, on average, are closer to the true values.

### Feature Importance Analysis

As evident in Figure 2.3, it is unsurprising that store size significantly influences the decision-making process of the Random Forest Regressor in predicting potential values.

A blue line with black text

Description automatically generated

*Figure 2.1: Contribution of features to the model’s decision-making process.*

## Conclusion

The FATER Geo-Analytics Challenge offers valuable insights for businesses in the Metropolitan City of Naples, particularly within the Italian diaper market. The analysis delves into socio-demographic factors, territorial features, and points of interest to uncover strategic considerations. Noteworthy findings include the dominance of hypermarkets and larger stores, the significance of parking availability, and the pronounced gravitational pull during weekends and specific time slots.

These insights paint a comprehensive picture of the retail landscape, providing a nuanced understanding of customer behavior and market dynamics. The strategic recommendations derived from the analysis encompass prioritizing hypermarkets, tailoring marketing efforts to specific demographics, and capitalizing on peak hours for promotions.

By prioritizing these actionable strategies, businesses can enhance their market positioning, cater to customer preferences, and optimize resource allocation. The comprehensive nature of the analysis ensures a strategic approach that aligns with the unique dynamics of the Italian diaper market in the Naples region.

For a detailed implementation plan and further tactical suggestions, the subsequent Recommendation section provides targeted guidance based on the findings outlined in this analysis.

## Findings and Observations:

* **With respect to existing customers:**

These observations can help with customer retention

* + Bigger stores tend to have more potential for drawing in customers and generating sales.
  + Stores with parking spaces tend to have higher potential values.
  + Hypermarkets tend to show higher potential values.
  + 80% of the combined potential is being generated by 20% (373) of the stores.
  + Hyperstores, which make up less than 20% of the store types, generate more than 80% of the combined potential.
* **With respect to new customers:**

These observations can help with customer acquisition

* + The population is likely to gravitate more during weekends.
  + Population is likely to gravitate more during 14 pm – 17 pm and 17 pm – 20 pm.
  + Population with age groups 41-50 and over 60 years are likely to gravitate more towards stores.
  + Males, on average, gravitate more towards stores.

## Recommendations

The strategies will contribute not only to the retention of existing customers but also to the acquisition of new customers, thereby increasing the store's potential for diaper products in Naples.

* **Strategic Marketing Focus:** 
  + Allocate marketing resources to prioritize hypermarkets and larger stores, given their higher potential.
  + Tailor marketing campaigns to emphasize the unique offerings of hypermarkets.
* **Peak Hour Promotions:**
  + Schedule promotions during peak evening hours (17:00–20:00) to capitalize on heightened customer gravitation.
* **Demographic Targeting:**
  + Design targeted campaigns for age groups 41–50 and those over 61, highlighting diaper options for both children and adults.
  + Craft marketing materials specifically designed to appeal to male customers, acknowledging their higher average gravitation towards stores. Highlight the features and benefits of diapers suitable for men, ensuring inclusivity in the product messaging.
* **Parking Facility Emphasis:**
  + Emphasize and promote stores with parking facilities to leverage the positive correlation with potential.

## Complimentary

### Population Dynamics up till 2043

A graph of different colored lines

Description automatically generated with medium confidence

*Figure 2.4: Trends of Live Births and Ageing Population up till 2043 in Naples.*

The trends depicted in Figure 2.4 clearly indicate a shift in the population landscape of Naples. While live births are declining, there is an apparent upward trend in the aging population, presenting an opportunity to increase sales of diapers for the elderly.

### Population and Household Projections for South Italy

* **Couples with Children:** A decline of more than seven percentage points in couples with children, from 36.1 percent in 2022 to 28.5 percent in 2042.
* **Couples with “Young” Children:** Couples with at least one child under 20 years are projected to decrease by about seven percentage points.
* **Single Persons:** Single persons are expected to grow more rapidly than the national average, reaching 35 percent in 2042, starting from 30.3 percent.
* **Childless Couples:** Childless couples will increase from 17.9 percent to 19.5 percent.
* **Lone Parents' Presence:** Lone parents' presence will increase to 13.3 percent in 2042.
* **Average Family Size:** The average family size is expected to decrease from 2.44 components in 2022 to 2.18 components in 2042.

### Strategies

* **Diversification of Product Line:** Introduce a wider range of diaper products catering to different age groups, including toddlers and preschoolers.
* **Bulk Discount:** Offer discounts for bulk purchases to appeal to caretakers and family members buying for elderly relatives.
* **Create a Referral Program:** Establish a referral program that incentivizes healthcare professionals to recommend your brand.
* **Urinary Incontinence Awareness Campaign:** A comprehensive TV, radio, and print campaign to destigmatize urinary incontinence, using engaging stories and expert interviews to encourage open conversations and promote Fater's commitment to community well-being.

## Bibliography

Alan Agresti, Maria Kateri. “*Foundations of Statistics for Data Scientists with R and Python*”. CRC Press. 2022.

Sebastian Raschka, Yuxi (Hayden) Liu, Vahid Mirjalili. “*Machine Learning with PyTorch and Scikit-Learn*”. Packt Publishing Ltd. 2022.

Aurélien Géron. *“Hands-On Machine Learning with Scikit-Learn & TensorFlow”.* O’Reille Media. 2019.

## Libraries & Modules

* pandas
* numpy
* seaborn as sns
* matplotlib
* statsmodels.api
* scipy.stats
* matplotlib.colors
* geopandas
* loads from shapely.wkt
* train\_test\_split, cross\_val\_score, KFold, GridSearchCV from sklearn.model\_selection
* RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor from sklearn.ensemble
* mean\_absolute\_error, r2\_score from sklearn.metrics

1. Ron Kohavi. *"A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection"*. International Joint Conference on Artificial Intelligence (IJCAI), 14 (2): 1137-43, 1995. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf [↑](#footnote-ref-1)