ALY 6980 Capstone Real Estate Agent Review Analysis

Professor – Valerie Atherley

Group 03

Mimansha Kaushik

Date – 12/04/2024

The problem: mismatched expectations

Traditional Approach

Traditional agent matching methods rely heavily on personal networks and referrals, often failing to address individual preferences and specific agent expertise.

The Result

This lack of personalization leads to mismatched expectations and less-than-ideal experiences for buyers, resulting in frustration and dissatisfaction.

Methodology

Data Sources

- Agent Data
- Agent Review
- Past Sales

Machine Learning Models

- Random Forest Regression, Linear Regression, Gradient Boosting Machines
- Classification Models, Support Vector Machines (SVM)

NLP Techniques

- NLP - Named Entity Recognition (NER)
- NLP - Sentiment Analysis

Recommendation

- Front-End Development for Recommendation System

Data processing

Data Cleaning:

- Missing value: Ratings, property prices were imputed using the median to minimize bias
- Extract State from address

Date Standardization:

- Filter past sales record from 2010 to most recent date to match the time range in agent review dataset

Data Merging:

- Use loc_id as common key to merge past sales and agent review
- Use name and location(state, pin code) to match agent information with agent review and past sales)

Data processing

Named entity recognition (NER):

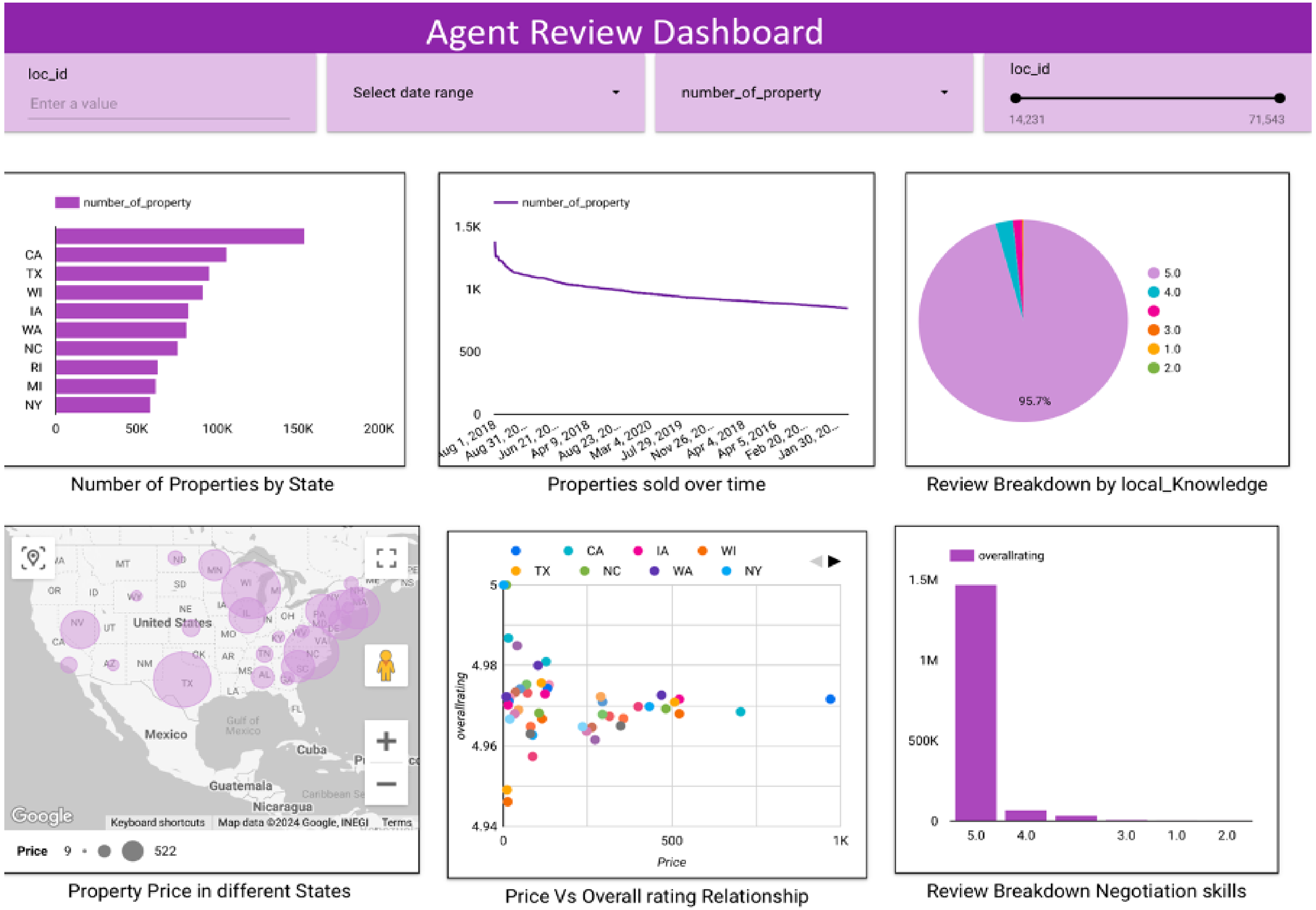
Package: spacy

Extract agent name from reviews from agent review, use it as potential key to make agent name in agent data. For reviews that may have multiple names, the most common name will be chosen as the identifier.

Sentiment analysis:

Package: textblob

Assess the polarity and subjectivity of customer reviews. A polarity greater than 0 will be categorized as positive feedback, while a polarity less than 0 will be considered negative feedback. This analysis can then be utilized to determine the percentage of positive comments for an agent in the subsequent step.



Agent Review Dashboard

loc_id
Enter a value

Select date range

number_of_property

loc_id
14,231 71,543

Number of Properties by State

Properties sold over time

Review Breakdown by local_Knowledge

Property Price in different States

Price Vs Overall rating Relationship

Review Breakdown Negotiation skills

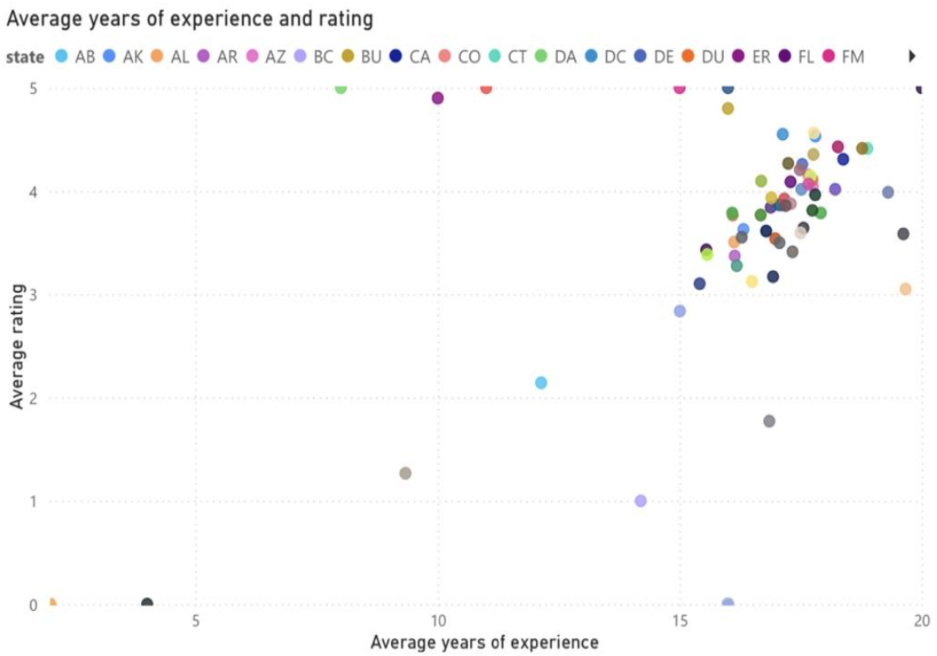
Looker dashboard

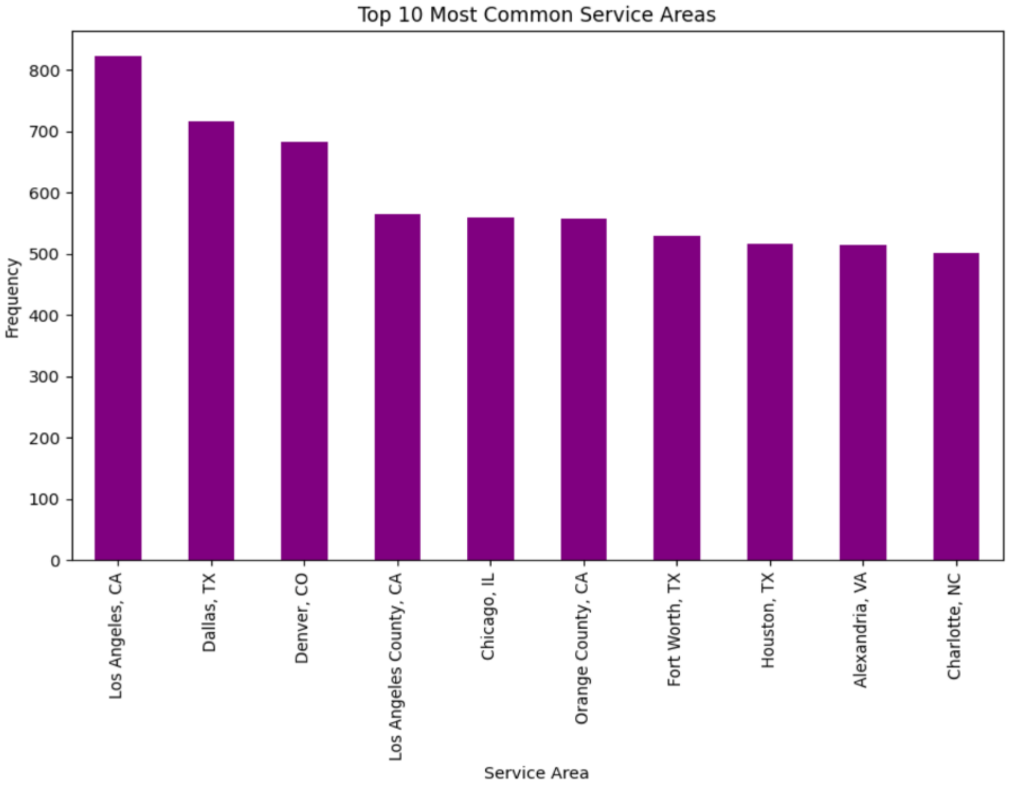
Exploratory data analysis (EDA)

Strong correlation between localized expertise and customer satisfaction, indicating that agents with deep knowledge of specific areas tend to receive higher ratings.

Agents with more experience consistently receive higher ratings, suggesting a strong relationship between experience and positive customer feedback.

Service areas with a high volume of transactions reflect regional demand, providing valuable insights into market trends and potential opportunities.





Model Building

Main Goal: Recommend agents based on customer preferences and expertise.

Models:

Regression: Predict agent performance scores.

Classification: Categorize agents into performance tiers.

Linear Regression:

1. As a baseline model, Linear Regression exhibited limited explanatory power, with only 10% of variance explained.
2. The high MSE indicates difficulty in capturing non-linear relationships.

Random Forest Regression:

1. Enhanced performance compared to Linear Regression by accounting for non-linear interactions.
2. The modest R^2 reflects challenges in fully capturing the complexity of agent suitability prediction.

Gradient Boosting Machines (GBM):

1. GBM emerged as the top-performing regression model, leveraging iterative improvement to reduce error and enhance explanatory power.
2. The reduced MSE and higher R^2 reflect better model fit and predictive accuracy.

Linear Regression:	Random Forest Regression:	Gradient Boosting Machines (GBM):
• **R^2:** 0.1048	• **R^2:** 0.2230	• **R^2:** 0.2878
• **MSE:** 3.5536	• **MSE:** 3.0845	• **MSE:** 2.8273

Model Building

Logistic Regression: **Accuracy:** 80.65%

Support Vector Machines (SVM): **Accuracy:** 84.64%

Logistic Regression Analysis:

- Simplistic yet effective, Logistic Regression achieved over 80% accuracy in binary classification (e.g., high vs. low suitability).
- Limited in capturing nuanced or multi-class suitability requirements.

Support Vector Machines Analysis:

- Outperformed Logistic Regression, with non-linear kernels effectively handling high-dimensional agent-customer data.
- Delivered improved accuracy by capturing complex relationships.

Achieving Our Goals

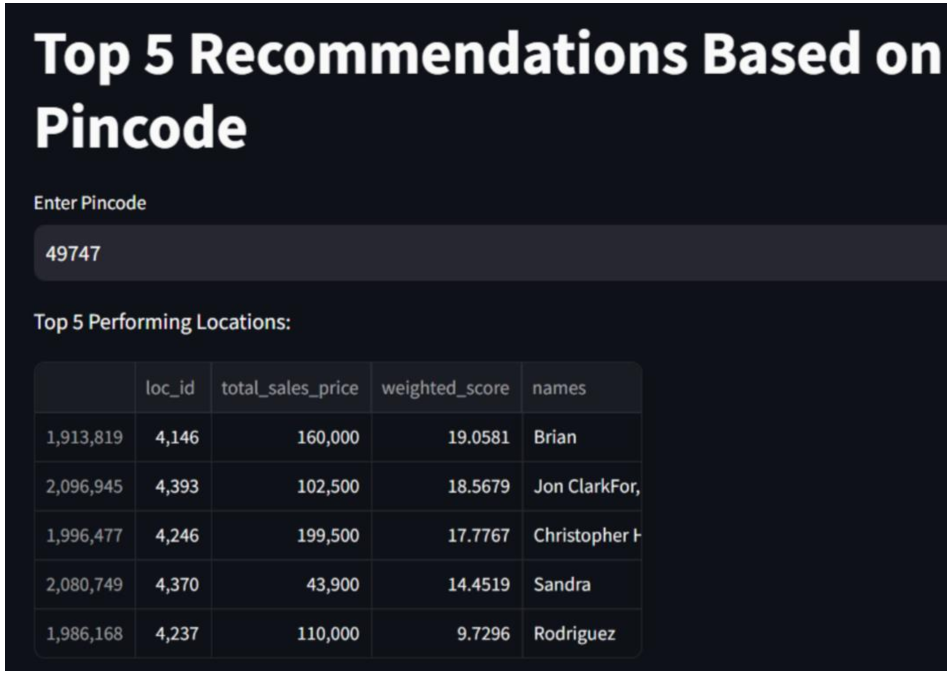
How We Achieved It:

- Engineered key features from profiles and reviews.
- Applied regression and classification models.
- Evaluated with R^2, MSE, and accuracy metrics.

Outcome:

- Accurate, data-driven agent recommendations for Buyer Folio.

Recommender system development



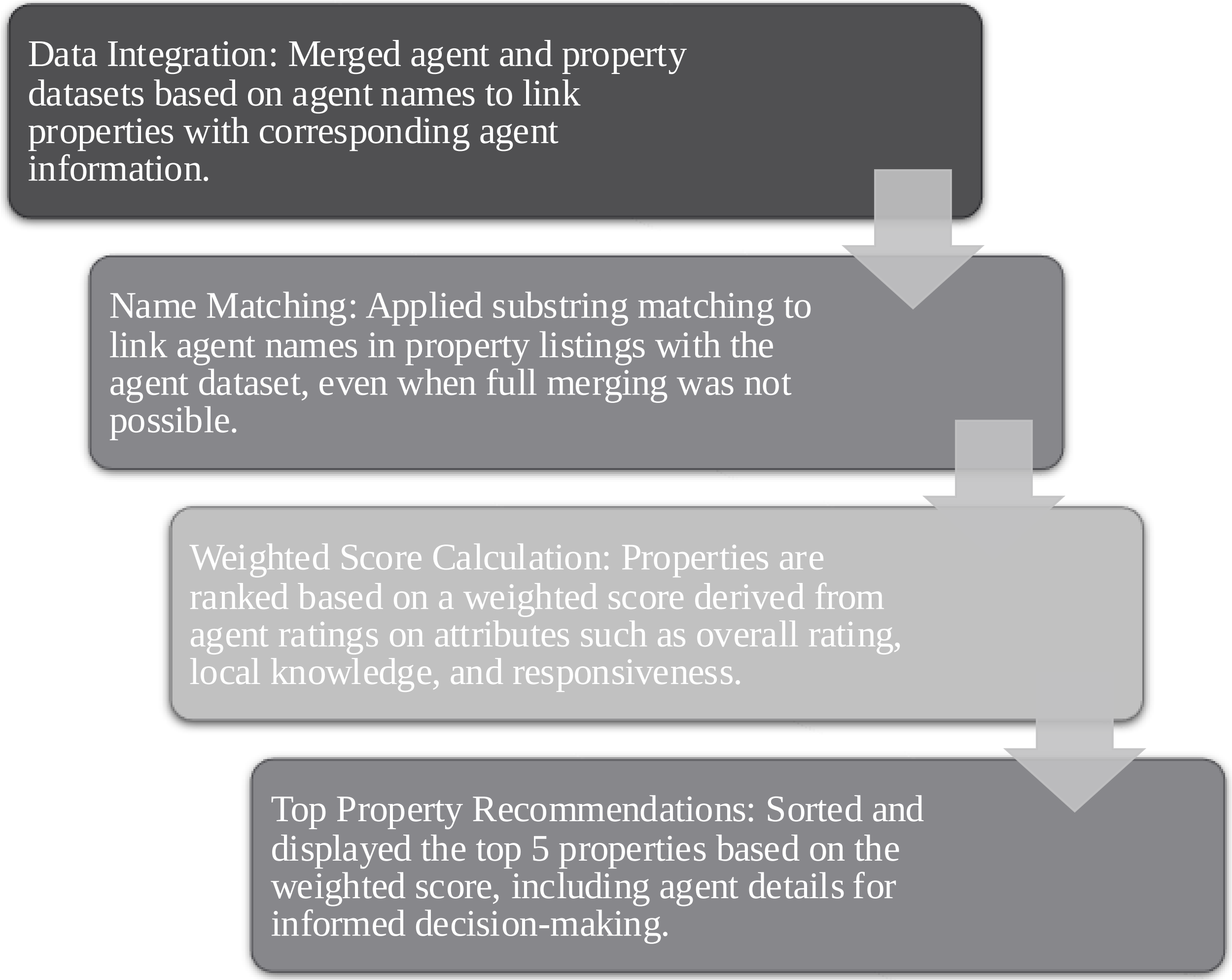
Top 5 Recommendations Based on Pincode

Enter Pincode

49747

Top 5 Performing Locations:

	loc_id	total_sales_price	weighted_score	names
1,913,819	4,146	160,000	19.0581	Brian
2,096,945	4,393	102,500	18.5679	Jon ClarkFor,
1,996,477	4,246	199,500	17.7767	Christopher H
2,080,749	4,370	43,900	14.4519	Sandra
1,986,168	4,237	110,000	9.7296	Rodriguez



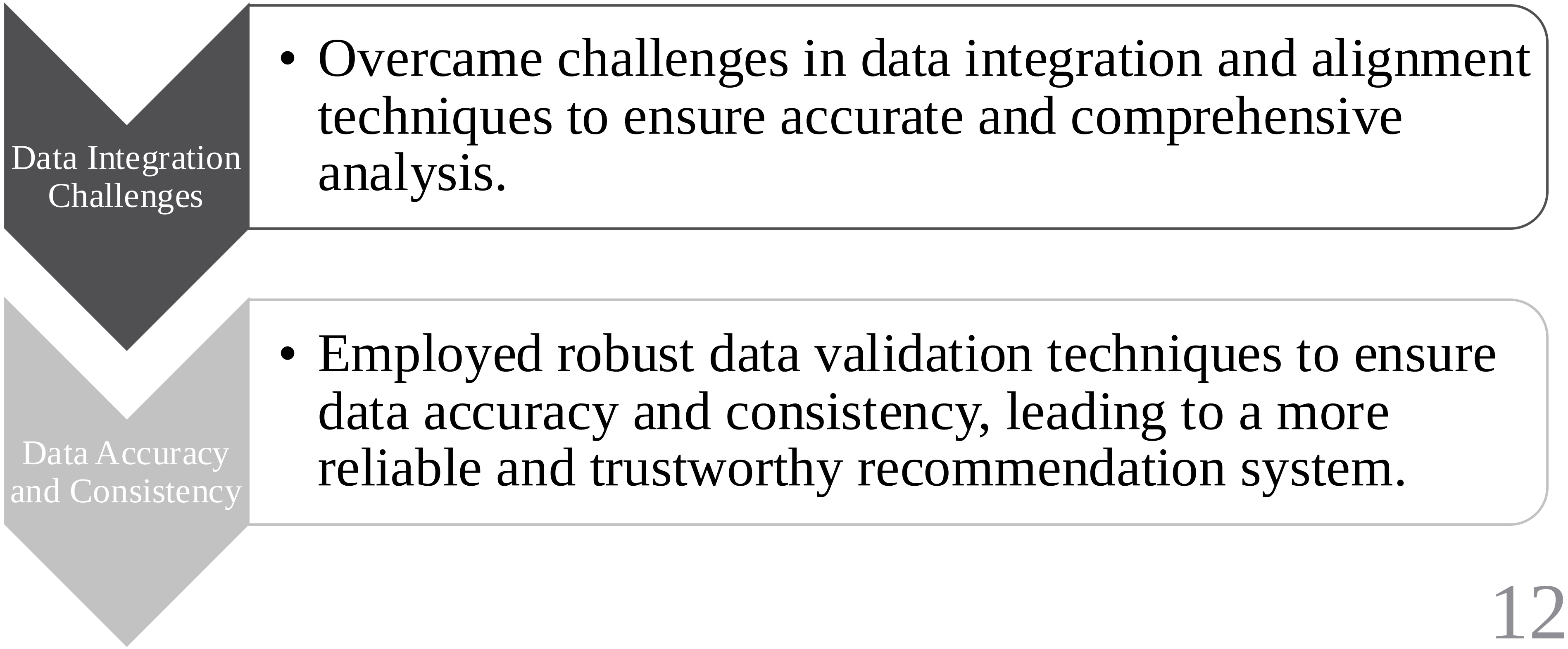
Data Integration: Merged agent and property datasets based on agent names to link properties with corresponding agent information.

Name Matching: Applied substring matching to link agent names in property listings with the agent dataset, even when full merging was not possible.

Weighted Score Calculation: Properties are ranked based on a weighted score derived from agent ratings on attributes such as overall rating, local knowledge, and responsiveness.

Top Property Recommendations: Sorted and displayed the top 5 properties based on the weighted score, including agent details for informed decision-making.

Challenges and solutions



Data Integration Challenges

- Overcame challenges in data integration and alignment techniques to ensure accurate and comprehensive analysis.

Data Accuracy and Consistency

- Employed robust data validation techniques to ensure data accuracy and consistency, leading to a more reliable and trustworthy recommendation system.

Key takeaways

This project has delivered a powerful data-driven solution that addresses the challenge of matching buyers with the most suitable agents. The system has achieved high accuracy and reduced error rates, ensuring personalized and effective recommendations for clients. This underscores the transformative potential of AI in the real estate industry and beyond.

References

- Yadav, Dinesh. “Categorical Encoding Using Label-Encoding and One-Hot-Encoder.” Medium, Towards Data Science, 9 Dec. 2019, towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd.
- Smith, J. A., & Doe, R. L. (2023). Machine learning applications in real estate recommendations. Journal of Real Estate Analytics, 15(3), 213-228.
- Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10), 4241-4251.
- Levitt, S. D., & Syverson, C. (2008). Market distortions when agents are better informed: The value of information in real estate transactions. *Review of Economics and Statistics*, 90(4), 599-611.

Thank You!