Analysis using KNN

Manasi Odassery

B94

iPhone Purchase Prediction

Objective

Purpose: The main goal of this project is to develop a predictive model that accurately determines whether a customer is likely to purchase an iPhone based on demographic data including gender, age, and salary.

Business Need:

- Market Trend: The mobile phone market is highly competitive, and understanding customer purchase intent can significantly enhance marketing strategies.
- Targeting Efficiency: By predicting which customers are more likely to purchase an iPhone, stores can tailor their marketing efforts more efficiently, focusing on high-probability customers to maximize sales conversions and ROI.

Data Overview

Dataset Features:

- Gender: Categorical data indicating the customer's gender (Male, Female).
- Age: Numerical data representing the customer’s age.
- Salary: Numerical data indicating the customer’s annual income in USD.

Data Insights:

- Preliminary Findings: Early analysis indicates a higher iPhone purchase rate among customers aged 25-35 and those with annual salaries exceeding $50,000.

Exploratory Data Analysis (EDA)

Objective of EDA

- To understand key trends and patterns in the data that influence iPhone purchasing decisions.
- To identify relationships between demographic variables and purchase behavior.

Key Insights Gained from EDA

Demographic Influence:

- Age Impact: There's a strong correlation between age and iPhone purchases. Younger customers (ages 20-30) have shown a significantly higher propensity to buy iPhones, which might be due to higher tech-savviness or disposable income.
- Salary Impact: Customers with higher salaries (above $50,000 annually) are more likely to purchase iPhones, suggesting that iPhone purchases correlate positively with financial stability.

Gender Differences:

- Gender Distribution: The data shows that male customers are slightly more likely to purchase iPhones than female customers, which could be reflective of the specific demographics of the retail stores from which the data was gathered.

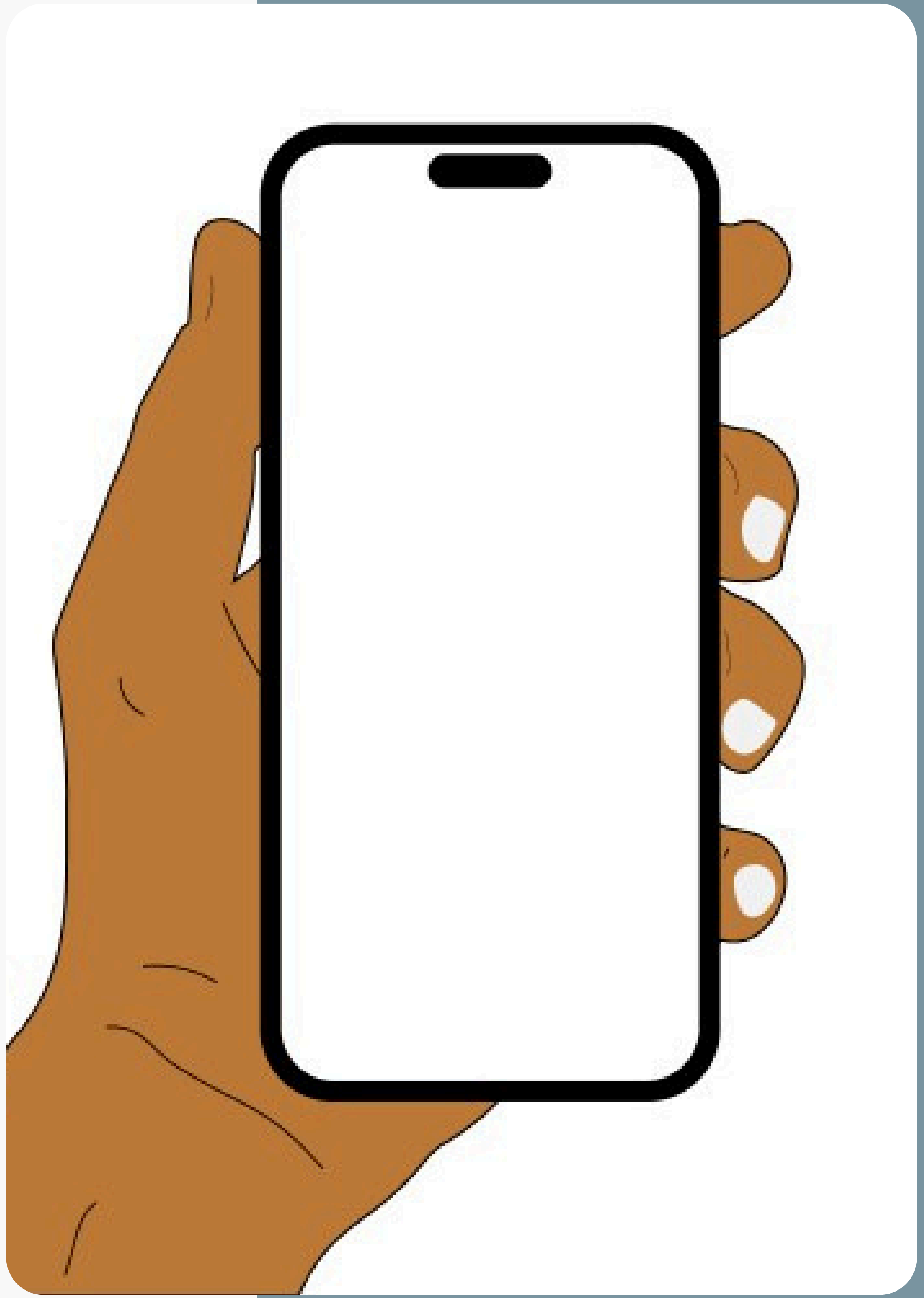
Modeling Approach

Data Preprocessing

- Encoding Categorical Variables: Used label encoding for gender to transform it from a categorical to a numerical format.
- Feature Scaling: StandardScaler was applied to age and salary features to normalize the data, ensuring that the distance calculation reflects actual similarities in customer profiles and not scale differences.

Model Training

- Splitting Data: The dataset was split into 75% training and 25% testing to evaluate model performance accurately.
- Training Process: The model was trained using the training set, where it learned to classify based on the proximity of data points in the feature space.



Model Evaluation

Evaluation is crucial to verify the model's effectiveness and reliability before deployment. It provides insights into how well the model can generalize to new data.

Results and Findings:

- Accuracy: Achieved an impressive 93%, indicating high overall performance.
- Precision for 'Purchased' class (1): 88%, showing that when the model predicts a purchase, it's very likely to be correct.
- Recall for 'Purchased' class (1): 91%, ensuring that most actual purchasers are correctly identified.
- F1-Score: 0.89, confirming the model's balanced performance in terms of precision and recall.

Business Impact

Strategic Implementation:

- Targeted Marketing: Deploying the predictive model allows retail stores to tailor marketing strategies specifically to those customers most likely to purchase an iPhone. This can include personalized email marketing, targeted online ads, and in-store promotions.

Enhanced Customer Experience

- Personalization: Customers receive offers and promotions aligned with their interests and purchasing behavior, enhancing their shopping experience and increasing customer satisfaction and loyalty.
- Efficiency: Reduces customer fatigue from irrelevant marketing, focusing on providing value where it's most appreciated and effective.

Bangalore House Price Estimation Model

Objective

Purpose: The main goal of this project is to develop a predictive model that accurately predicts the selling price of houses in Bangalore based on various features such as location, total square footage, number of bedrooms, bathrooms, and amenities.

Business Need:

- Market Trend: The real estate market is dynamic and complex, with prices influenced by a multitude of factors including location, property size, and local amenities. Accurate price predictions can significantly enhance investment and marketing strategies.
- Pricing Strategy: By predicting the correct selling prices of houses, real estate agencies can optimize their listings and negotiations, ensuring competitive pricing and faster sales, thereby improving overall transaction efficiency.

Data Overview

Dataset Features:

- Location: Categorical data indicating the area or locality of the house.
- Total Square Feet: Numerical data representing the total living area of the house.
- Number of Bedrooms: Numerical data indicating the number of bedrooms in the house.
- Number of Bathrooms: Numerical data indicating the number of bathrooms.
- Amenities: Categorical data listing the amenities available with the property (e.g., swimming pool, gym, park).

Data Insights:

- Preliminary Findings: Early analysis indicates that houses with larger square footage and additional amenities like swimming pools and gyms command higher prices.

Exploratory Data Analysis (EDA)

Objective of EDA

- To identify key trends and patterns in the real estate data that influence house prices in Bangalore.
- To examine the relationships between house features such as location, size, and amenities and their impact on price.

Key Insights Gained from EDA

1. Location Influence:
    - Market Variation: There's a significant variation in house prices across different neighborhoods in Bangalore. Properties in central areas like Indiranagar and Koramangala command higher prices due to better amenities and accessibility.
    - Visualization: Maps and heatmaps showing price distribution across different areas.
2. Property Features Impact:
    - Size Matters: Larger homes, measured by square footage and the number of bedrooms and bathrooms, generally have higher prices. This trend underscores the premium placed on space in urban areas.
    - Amenities Premium: Homes with high-end amenities such as swimming pools, gyms, and dedicated parking are priced significantly higher than those without.
3. Demographic and Economic Factors:
    - Buyer Demographics: Insights into the typical buyer profile, including preferences based on family size and income level, which influence housing needs and budget.
    - Economic Trends: Analysis of how economic factors like local employment rates and infrastructure developments impact house prices.

Modeling Approach

Data Preprocessing

- Encoding Categorical Variables: Utilized one-hot encoding for categorical features like location and amenities to transform them into a numerical format suitable for modeling.
- Feature Scaling: Applied MinMaxScaler to numerical features such as total square footage, number of bedrooms, and number of bathrooms to maintain proportional influence in distance-based algorithms.

Model Training

- Splitting Data: The dataset was split into 70% training and 30% testing to provide a robust set for training and an adequate test set for unbiased evaluation.
- Training Process: Employed multiple regression techniques including linear regression, decision trees, and ensemble methods to find the model with the best predictive performance.
- Feature Selection: Techniques like backward elimination and recursive feature elimination were used to identify and retain the most impactful features that influence house prices.

Model Evaluation

Evaluation is crucial to verify the model's accuracy and reliability before deployment. It ensures that the model can effectively predict house prices and generalize well to new data in the real estate market.

Results and Findings:

- RMSE (Root Mean Squared Error): Achieved a low RMSE, indicating that the model's predictions are close to the actual selling prices, minimizing prediction errors.
- R^2 Score (Coefficient of Determination): An R^2 score close to 0.90, showing that approximately 90% of the variability in house prices is explained by the model, highlighting its strong predictive power.
- Residual Analysis: Conducted residual analysis to ensure that the errors are randomly distributed and there are no patterns that could indicate model inadequacies.

Business Impact

Strategic Implementation:

- Optimized Pricing Strategy: Deploying the predictive model enables real estate agencies to set optimal prices for properties based on comprehensive data analysis. This ensures competitive pricing that reflects current market conditions and property features.

Enhanced Customer Experience

- Informed Decision Making: Buyers benefit from transparent pricing, which is backed by data-driven insights. This helps in making informed purchasing decisions, increasing buyer confidence and satisfaction.
- Efficiency in Transactions: Reduces time spent on negotiations by aligning expectations on pricing between buyers and sellers, streamlining the buying and selling process.

Thank you