ANALYSIS USING RANDOM FOREST

MANASI ODASSERY
B94

SALES CLASSIFICATION ANALYSIS USING RANDOM FOREST

OBJECTIVE

- Purpose: The primary goal of this project is to understand the key drivers that influence sales outcomes at different store locations. By identifying these factors, the company can better strategize its marketing efforts, optimize product placements, and tailor its pricing strategies to maximize sales.
- Analytical Approach: Utilizing a Random Forest classification model allows us to explore the complex interactions between multiple variables such as competitor prices, local income levels, advertising expenditures, and more. This model will help us categorize store locations into distinct sales performance categories: Low, Medium, and High.

DATA OVERVIEW

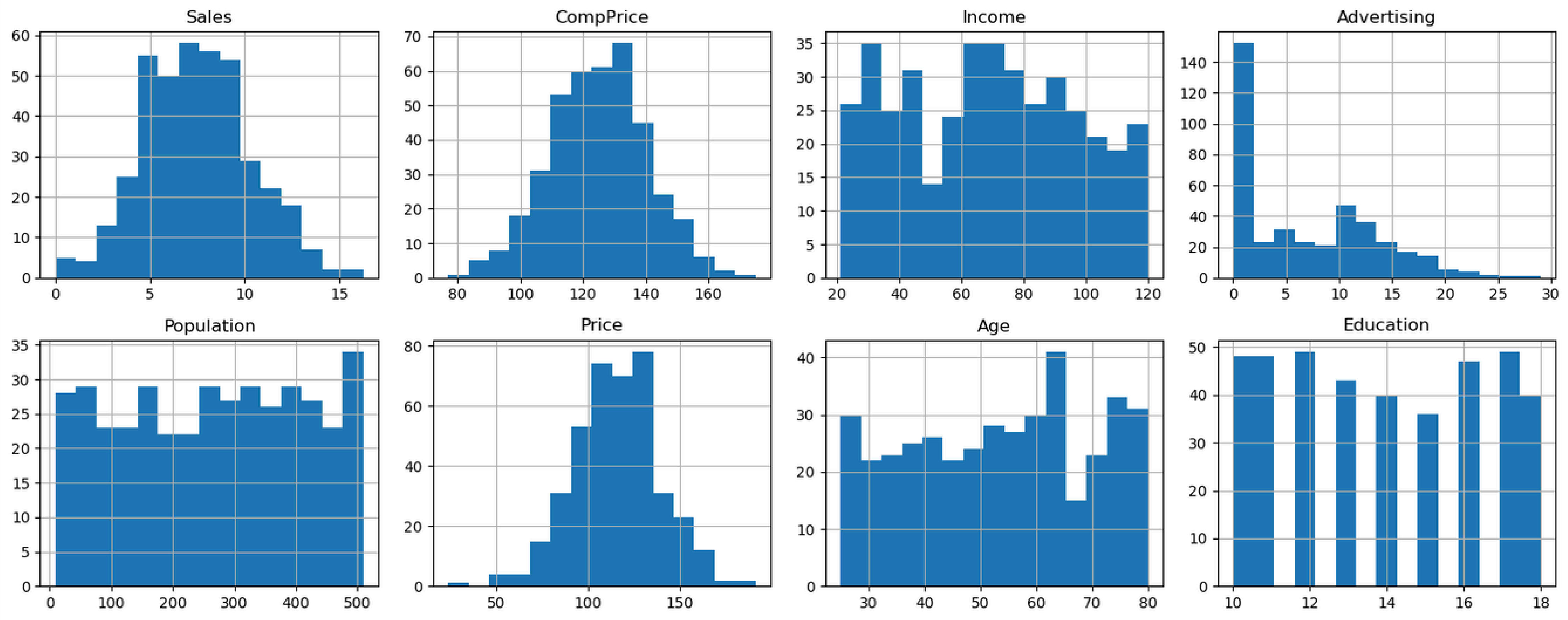
Data Description: The dataset for this project consists of 400 entries, each representing a store location. The variables include:

- Sales: Unit sales at each location (in thousands).
- Competitor Price: Price charged by competitors at each location.
- Income: Average community income level at each location (in thousands of dollars).
- Advertising: Advertising budget for the company at each location (in thousands of dollars).
- Population: Population size in the region (in thousands).
- Price: Price charged by the company for car seats at each site.
- Shelf Location: Quality of shelving location (Bad, Good, Medium).
- Age: Average age of the local population.
- Education: Education level at each location.
- Urban: Whether the store is in an urban or rural area.
- US: Whether the store is located in the US.

EXPLORATORY DATA ANALYSIS

Key Findings:

- Sales Distribution: Sales are heavily skewed towards lower volumes, suggesting high competition in most locations.
- Impact of Competitor Prices: Higher competitor prices are associated with increased sales at our locations, indicating a strong position in markets where competitors charge more.
- Income and Sales: Higher community income levels correlate positively with sales, highlighting income as a key factor in purchasing decisions.
- Advertising Spend: The effectiveness of advertising varies significantly across regions, with no clear pattern suggesting a one-size-fits-all approach may not be effective.



METHODOLOGY

Data Preprocessing

- Categorized sales into three levels: 'Low', 'Medium', and 'High' based on quantiles to enable classification.
- Normalized numeric variables like 'Income' and 'Advertising' to ensure they are on a similar scale, improving model performance.

Model Training and Validation

- Split the data into training (70%) and testing (30%) sets to evaluate the model's performance on unseen data.
- Employed cross-validation during training to ensure the model's stability and to prevent overfitting.

Model Selection

- Random Forest Classifier is known for its ability to handle a large number of features and its robustness against overfitting.
- Utilized hyperparameter tuning through Grid Search to find the optimal settings for parameters like the number of trees and depth of the trees.

Evaluation Metrics

- Evaluated the model using accuracy, precision, recall, and F1-score to assess its performance across all sales categories.

MODEL PERFORMANCE

Performance:

- The Random Forest model achieved an accuracy of 65.83%, demonstrating a solid capability to classify sales levels.

Class-Specific Performance:

- High Sales: Achieved a precision of 71% and a recall of 73%, indicating the model is fairly reliable at identifying high sales locations.
- Medium Sales: With precision at 58% and recall at 56%, this category shows room for improvement in prediction accuracy.
- Low Sales: Precision and recall both around 68%, showing balanced performance but potential for enhancement.

Analysis:

- The model performs best in distinguishing high sales locations, which is critical for targeting high-potential markets.
- Medium sales are the hardest to predict, possibly due to overlapping characteristics with other categories.

BUSINESS IMPACT

Strategic Decision-Making

The model's ability to accurately classify sales levels allows for targeted marketing strategies. By understanding which locations are likely to generate high, medium, or low sales, the company can allocate resources more effectively.

Inventory Management

Predictive insights from the model enable better stock level management. High sales locations can stock more inventory to meet demand, whereas low sales areas can reduce overstock to minimize losses.

Resource Optimization

Advertising budgets can be adjusted based on the predicted sales impact. Locations with high sales potential might receive increased advertising spend, while areas with consistently low sales could see reduced expenditure or different marketing tactics.

Revenue Growth

By optimizing marketing and inventory strategies based on model predictions, the company can potentially increase revenue at high-potential locations and decrease unnecessary costs in lower-performing areas.

FRAUD DETECTION ANALYSIS USING RANDOM FOREST

OBJECTIVE

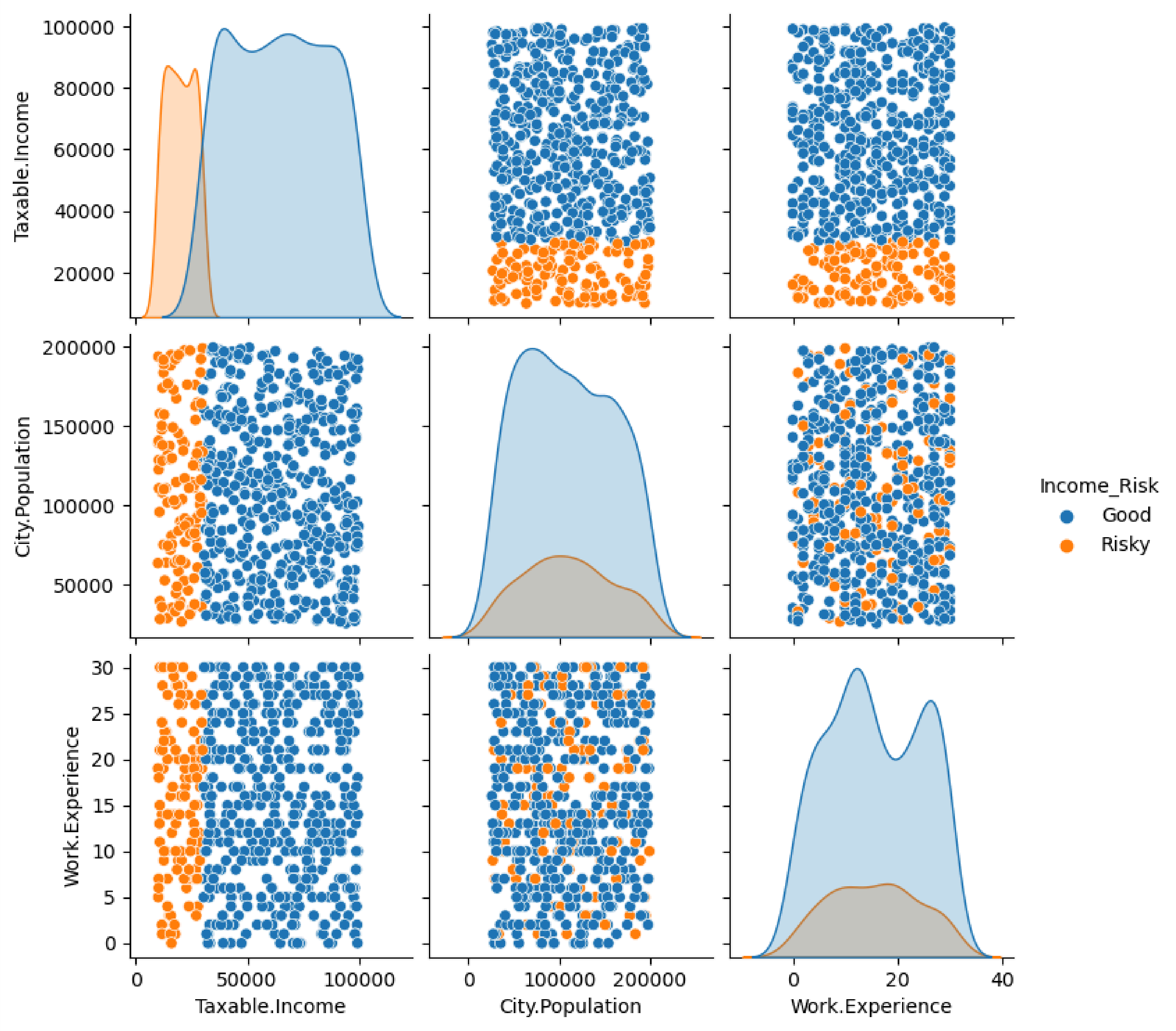
- Purpose: This project is aimed at developing a predictive model to classify individuals based on their risk of engaging in fraudulent financial behavior. By focusing on taxable income and other relevant financial indicators, we can proactively identify high-risk profiles.
- Analytical Approach: The Random Forest Classifier is chosen for its robust performance in classification tasks and its ability to handle imbalanced datasets, which is crucial given the uneven distribution of 'Risky' versus 'Good' classifications in our dataset.
- Business Relevance: Effective fraud detection is critical for maintaining financial security and integrity within the company. This model will facilitate early intervention, potentially saving substantial costs associated with fraud investigations and losses.

DATA OVERVIEW

Data Description: The fraud detection dataset includes 600 records, with each entry detailing financial and demographic information of an individual. Key variables include:

- Undergrad: Whether the individual has a college degree.
- Marital Status: Marital status of the individual.
- Taxable Income: Annual taxable income.
- City Population: Population of the city where the individual resides.
- Work Experience: Number of years of work experience.
- Urban: Whether the individual lives in an urban area.

EXPLORATORY DATA ANALYSIS



Key Observations:

- Income Risk Categories: Defining 'Risky' as taxable income <= 30,000 revealed that a significant minority of the dataset falls into this high-risk category.
- Demographic Influences: Marital status and urban residency have notable correlations with income levels, potentially influencing the risk of fraud.
- Work Experience: Surprisingly, there is no clear trend between years of work experience and income levels, suggesting that fraud risk is less dependent on professional experience.

METHODOLOGY

Data Preprocessing

- Defined 'Risky' based on individuals with taxable income <= 30,000 as per the project's criteria.
- Converted categorical features like 'Undergrad', 'Marital.Status', and 'Urban' into numeric codes using Label Encoding.

Model Training and Validation

- Data was divided into 70% training and 30% testing segments to validate the effectiveness of the model.
- Used stratified splits in cross-validation to ensure representative distribution of both classes in each fold.

Model Selection

- The Random Forest Classifier is known for its efficacy in classification tasks and its feature importance capabilities, which are crucial for understanding risk factors.
- Configured the model with parameters aimed at maximizing the detection of the minority class without losing overall accuracy.

Evaluation Metrics

- Performance measured by precision, recall, F1-score for each class, and overall accuracy to provide a comprehensive view of the model's capabilities.

MODEL PERFORMANCE

Performance:

- The Random Forest model reached an overall accuracy of 75.56%, which is commendable given the challenge of predicting fraudulent behavior.

Class-Specific Performance:

- Good: High precision (79%) and recall (95%), indicating effective identification of low-risk individuals.
- Risky: Both precision and recall are 0%, highlighting significant challenges in detecting high-risk cases.

Analysis:

- The model excels in identifying 'Good' class but fails to detect the 'Risky' class, likely due to class imbalance and the nature of the data.

BUSINESS IMPACT

Fraud Prevention

Early detection of potential fraud risks through the model allows the company to take preemptive measures, reducing the incidence of actual fraud and associated costs.

Risk Management

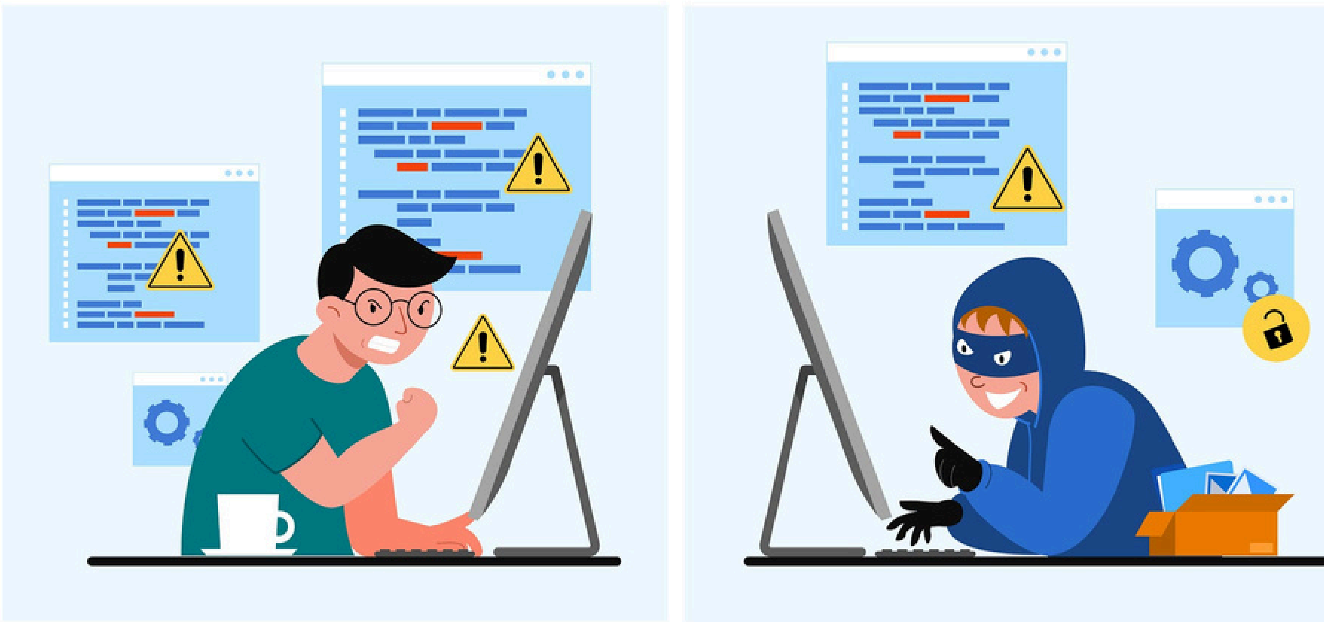
By identifying high-risk profiles, the company can tailor its customer monitoring and review processes, focusing on areas with the highest likelihood of fraud.

Operational Efficiency

Automating the initial risk assessment process with the model reduces the workload on human analysts, allowing them to focus on more complex investigations and customer interactions.

Customer Trust and Compliance

Effective fraud detection strengthens customer trust and enhances compliance with financial regulations, protecting the company's reputation and legal standing.



THANK YOU