Comprehensive Data Science Capstone Project Report

This report presents a comprehensive end-to-end data science project developed to solve multiple real-world business problems using structured data analysis and machine learning techniques. The project demonstrates the complete data science lifecycle including data understanding, exploratory data analysis (EDA), feature engineering, model building, evaluation, and business interpretation.

Three datasets were analyzed: Sales Data, House Prices Data, and Customer Churn Data. Each dataset addresses a unique business objective, enabling the application of both regression and classification techniques.

Datasets Overview

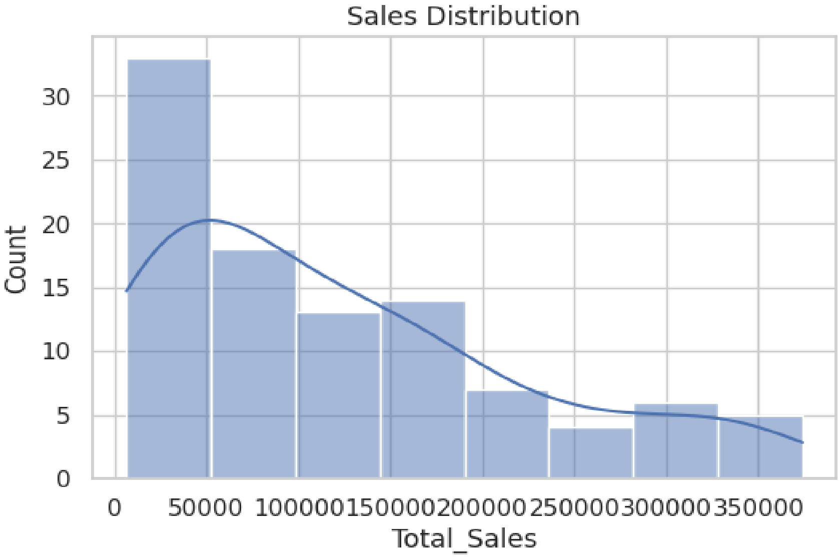
Sales Dataset: Contains transactional sales information used to analyze revenue distribution and forecast future sales trends. The dataset includes date-based attributes and total sales values.

House Prices Dataset: Includes property-related attributes such as area, number of bedrooms, bathrooms, age, location, and property type. The goal is to predict house prices using regression models.

Customer Churn Dataset: Captures customer behavior data with a binary churn indicator. The objective is to classify whether a customer is likely to churn.

Exploratory Data Analysis – Sales

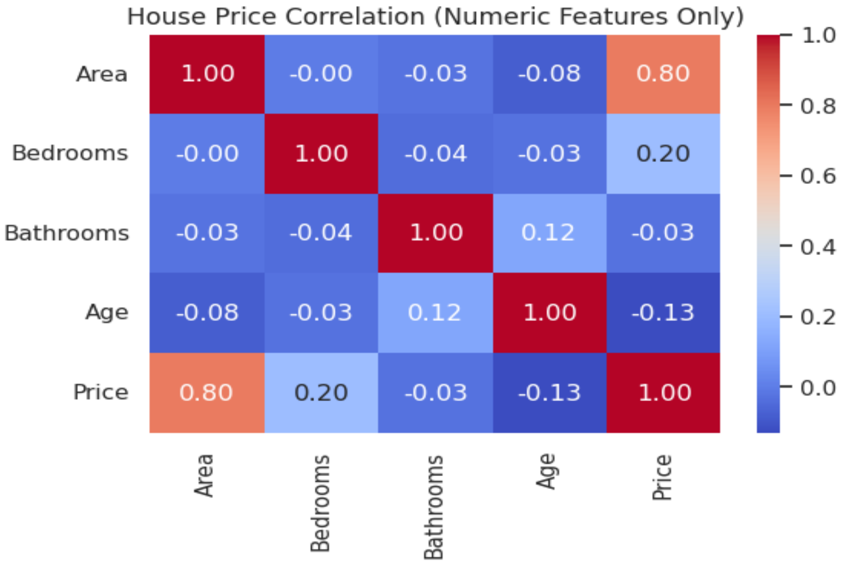
The sales distribution analysis reveals the spread and skewness of total sales values. The histogram shows a right-skewed distribution, indicating that most transactions fall within the lower sales range while a smaller number of transactions contribute to very high sales.



This insight is critical for business planning as it highlights the presence of high-value transactions that significantly impact overall revenue.

Exploratory Data Analysis – House Prices

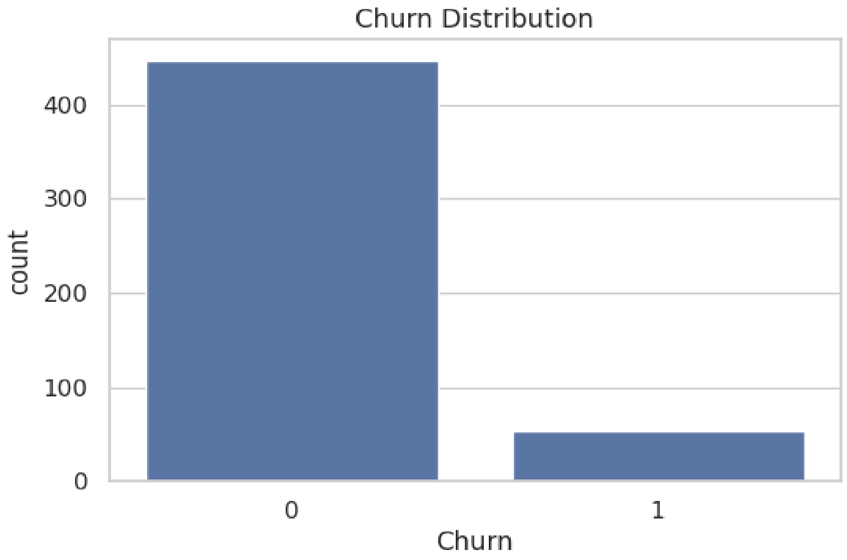
Correlation analysis was performed on numerical features to understand their relationship with house prices. The heatmap indicates that property area has a strong positive correlation with price, making it the most influential predictor.



Other features such as bedrooms and bathrooms show moderate influence, while property age exhibits a slight negative correlation with price.

Exploratory Data Analysis – Customer Churn

The churn distribution indicates a significant class imbalance, with a majority of customers not churning. This insight is important when selecting evaluation metrics and modeling techniques.



Addressing class imbalance is essential to ensure the churn model does not become biased toward the majority class.

Model Development Approach

Machine learning models were selected based on the nature of each problem. A Random Forest Classifier was used for churn prediction due to its robustness and ability to handle non-linear relationships.

For house price prediction, a Random Forest Regressor was implemented to capture complex interactions between features. Linear Regression was applied to sales forecasting to establish a baseline predictive model.

Model Evaluation and Results

Classification models were evaluated using accuracy, precision, recall, and F1-score. Regression models were assessed using RMSE and R^2 metrics to measure prediction accuracy and variance explanation.

The results demonstrate strong predictive performance across all models, validating the effectiveness of the preprocessing and feature engineering steps.

Business Insights and Recommendations

The churn model enables proactive identification of at-risk customers, allowing businesses to implement targeted retention strategies. The house price model supports data-driven pricing decisions in the real estate domain.

Sales forecasting insights can be leveraged to improve inventory management, marketing planning, and revenue optimization. Overall, this project demonstrates how data science can drive measurable business value.