Business Understanding

E6: Apartment Price Prediction and Analysis
Team Members: Robin Henrik Neem, Marten Ojasaar, Markus Nurmik
Repository: https://github.com/RobinHenrik/RealEstatePricePredicing

Task 2

1. Identifying Your Business Goals

Background
Understanding apartment prices and predicting their values is critical for buyers, sellers, and real estate investors. Accurately predicting apartment prices helps identify market trends, provide investment insights, and ensure fair transactions. This project aims to provide an advanced predictive model using data scraped from kv.ee, analyzing key features that influence apartment prices in Estonia.

Business Goals
1. Predict the sale price of properties based on features like location, size, number of rooms, and condition.
2. Identify the most influential factors affecting property prices to provide actionable insights for stakeholders.
3. Detecting undervalued properties to aid real estate investors in finding profitable opportunities.

Business Success Criteria
1. A predictive model achieving high accuracy, measured by metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), when applied to the test dataset.
2. Insights into key features affecting property prices, validated through feature importance analysis or statistical testing.
3. Identification of undervalued properties compared to market trends, demonstrated by actionable examples that investors find useful.

2. Assessing Your Situation

Inventory of Resources
1. Data: Scraped dataset from kv.ee with columns such as property ID, location (county, city, district), size, rooms, year built, condition, energy efficiency, ownership type, building material, and price.
2. Tools: Python, Jupyter Notebook, libraries for data processing (e.g., pandas, numpy), machine learning (e.g., scikit-learn, XGBoost), and visualization (e.g., matplotlib, seaborn).

3. Expertise: Team members possess skills in data scraping, machine learning, and statistical analysis.
4. Infrastructure: Local computational resources for model training and testing.

Requirements, Assumptions, and Constraints

1. Requirements:
    - Clean and structured data from kv.ee for robust analysis.
    - A feature-rich dataset with minimal missing or inconsistent values.
    - An evaluation framework for comparing model performance.
2. Assumptions
    - Market trends in Estonia remain stable during the data collection period.
    - Features like location, size, and condition are significant predictors of price.
    - Scraped data represents a wide range of properties in Estonia.
3. Constraints:
    - Limited to the scraped dataset, which may exclude certain regions or property types.
    - Computational resources may limit the complexity of the models used.

Risks and Contingencies

1. Risk: Insufficient or unrepresentative data may lead to inaccurate predictions.
   Mitigation: Use additional data sources if required and perform rigorous data validation.
2. Risk: Overfitting the model to the training dataset.
   Mitigation: Apply cross-validation techniques and regularization methods.
3. Risk: Model interpretability vs. complexity trade-offs.
   Mitigation: Use a combination of interpretable and complex models to balance accuracy and insights.

Terminology

- MAE: Mean Absolute Error, a measure of prediction accuracy.
- RMSE: Root Mean Squared Error, another measure of accuracy that penalizes larger errors.
- Feature Importance: A ranking of input features based on their impact on the prediction model.

Costs and Benefits

1. Costs:
    - Time investment for data cleaning, modeling, and evaluation.
2. Benefits:
    - Improved decision-making for buyers, sellers, and investors.
    - Insights into the real estate market, aiding policy and investment strategies.
    - Identification of undervalued properties, creating opportunities for profit.

3. Defining Your Data-Mining Goals

Data-Mining Goals

1. Develop a machine learning model capable of accurately predicting apartment prices.
2. Perform exploratory data analysis (EDA) to understand the distribution of property prices and features.
3. Conduct feature importance analysis to determine which attributes most significantly influence apartment prices.
4. Create a method to identify undervalued properties based on predicted versus listed prices.

Data-Mining Success Criteria

1. A predictive model achieving a target MAE or RMSE, ideally within 10% of industry benchmarks for real estate.
2. A clear set of influential features supported by statistical metrics (e.g., correlation, p-values).
3. Identification of at least 5 properties in the dataset as potentially undervalued, validated through manual or market comparisons.

Task 3

1. Gathering Data

Outline Data Requirements:
The dataset for this project must include key features essential for apartment price prediction. These features are:

- Location: Including county, city, and district, as real estate prices vary significantly by location.
- Size: Total area in square meters, as it directly correlates with price.
- Rooms: The number of total rooms and bedrooms, which are critical for valuation.
- Condition and Year Built: Reflecting the property's age and state, affecting its market value.
- Energy Efficiency: A growing concern for buyers, influencing desirability and price.
- Building Material and Ownership Type: Indicators of durability, style, and legal implications.
- Price: The target variable for prediction, representing the sale price of the property.

Verify Data Availability:
The primary data source is kv.ee, from which data was scraped successfully. The dataset contains critical columns including id, maakond (county), linn (city), linnaosa (district), pind (size), tube (rooms), magamistube (bedrooms), korrus (floor), korruseid (total floors), ehitusaasta (year built), seisukord (condition), energiamärgis (energy efficiency), hoone materjal (building material), omandivorm (ownership type), and hind (price).

Define Selection Criteria:
The dataset must include properties from diverse regions and price ranges in Estonia to ensure representativeness. Properties with missing critical data (e.g., size or price) or obvious errors will be excluded. The dataset will prioritize urban areas, where the real estate market is most active, but will not exclude rural properties.

2. Describing Data

The dataset contains 14 columns, each representing a specific attribute of a property. The key attributes include:

- Location Fields: maakond, linn, and linnaosa provide hierarchical location data.
- Property Characteristics: pind (size), tube (rooms), magamistube (bedrooms), korrus (floor), and korruseid (total floors).
- Condition Indicators: ehitusaasta (year built), seisukord (condition), energiamärgis (energy efficiency).
- Additional Features: hoone materjal (building material), omandivorm (ownership type).
- Target Variable: hind (price).

The data covers a wide variety of apartments, providing sufficient diversity for meaningful analysis.

3. Exploring Data

Initial exploration reveals:

- Distribution of Prices: Prices range from low-cost apartments to luxury properties, showing a right-skewed distribution.
- Size and Rooms: Size (in square meters) and the number of rooms exhibit a positive correlation with price. Outliers, such as unusually high prices for small properties, may indicate data entry errors or special cases (e.g., historic value).
- Location Impact: Urban properties, especially in Tallinn, generally show higher prices compared to rural areas.
- Missing Data: A few properties lack values for fields like energiamärgis or seisukord. Missing data will need handling during preparation.
- Feature Relationships: Features such as ehitusaasta and seisukord indicate a relationship where newer or well-maintained properties command higher prices.

4. Verifying Data Quality

Completeness: Most critical fields (e.g., pind, tube, hind) are well-populated. However, non-critical fields such as energiamärgis have some missing values.

Accuracy: Data scraping occasionally misclassifies features. For example, some properties list impossible combinations like a size of 0 square meters.

Consistency: Location fields show consistent formats, but some entries have spelling variations or abbreviations. Standardizing these is necessary.

Outliers: A few extreme outliers in the hind column (e.g., properties listed at 1 EUR or extremely high prices) are likely errors or test cases. These will be reviewed and addressed.

Action Plan for Quality Issues:

- Standardize location names and formats.
- Impute or exclude records with missing or nonsensical values.
- Investigate and handle outliers, ensuring they do not unduly influence the model.

By gathering, describing, exploring, and verifying the data, we have ensured a thorough understanding of its scope, limitations, and potential. This sets the foundation for effective data preparation and modeling in subsequent steps.

Task 4

1. Project plan

Task	Description	Hours/Member Contribution
Data Collection and Cleaning	Ensure the scraped dataset is complete, clean, and structured. Remove duplicates, handle missing values, and standardize feature formats.	Robin: 10 hrs, Marten: 12 hrs, Markus: 8 hrs
Exploratory Data Analysis (EDA)	Visualize the data to identify trends, outliers, and relationships between features and the target variable (price).	Robin: 6 hrs, Marten: 8 hrs, Markus
Feature Engineering and Selection	Select the most predictive features using statistical and machine learning	Robin: 8 hrs, Marten: 6 hrs, Markus: 8 hrs

	techniques.	
Model Development and Training	Train multiple models (e.g., linear regression, XGBoost) to predict apartment prices. Perform hyperparameter tuning for optimal performance.	Robin: 8 hrs, Marten: 10 hrs, Markus: 10 hrs
Results Analysis and Presentation	Analyze model performance, summarize findings, and create a presentation highlighting key insights, predictions, and recommendations for stakeholders.	Robin: 6 hrs, Marten: 8 hrs, Markus: 6 hrs

2. Methods and tools

- Data Collection and Cleaning: Python, pandas, NumPy.
- EDA: matplotlib, seaborn, Python.
- Feature Engineering and Selection: scikit-learn, statistical tests, feature importance analysis.
- Model Development: scikit-learn, XGBoost.
- Results Presentation: matplotlib, PowerPoint, Jupyter Notebook for visualizations and explanations.

Comments:

- Effective collaboration is critical; tasks will overlap, with team members reviewing each other's work.
- A detailed timeline will ensure we meet all deadlines, particularly for model development and presentation preparation.
- Regular check-ins will help identify issues early and maintain steady progress.