For advanced regression techniques, you could consider using ensemble methods which often perform well on structured data like housing prices:

- Random Forests
- Gradient Boosting Machines

As for feature engineering, techniques like domain-specific knowledge, interaction terms, or even automated feature selection methods could be beneficial.

What Are Interaction Terms in Feature Engineering?

Interaction terms are **new features created by combining two or more existing features to capture how they interact with each other**. They are especially useful when the effect of one feature on the target depends on the value of another feature.

Why Are They Important?

- In many real-world cases, the relationship between features and the target is **not purely additive** (meaning the effect of one feature might depend on another).
- Standard linear models assume all features act independently, which can miss these complex relationships.
- Interaction terms allow the model to **capture these dependencies**.

Hypothesis

With the House Prices dataset, you can explore several hypotheses related to housing prices and their determinants. Here are a few ideas:

1. **Location Influence**: Test whether housing prices vary significantly based on location factors such as neighbourhood, proximity to amenities, or school districts.
2. **Feature Importance**: Identify which features (like number of bedrooms, area size, etc.) correlate strongly with housing prices.
3. **Temporal Trends**: Investigate if there are temporal trends in housing prices over the years covered by the dataset.

Regarding professional, ethical, social, and sustainability issues:

- **Ethical Considerations**: Ensure your analysis respects privacy and data confidentiality, especially if the dataset includes sensitive information about homeowners.
- **Social Implications**: Consider how your findings might impact housing policies or community development strategies.
- **Sustainability**: Explore how features like energy efficiency ratings or green certifications impact housing prices, promoting sustainable housing practices.
-

Papers

"A Comparative Study of Machine Learning Algorithms for House Price Prediction" by Khaled Mohammed Saafan et al. This paper compares various machine learning algorithms for predicting house prices, which could provide a good benchmark for your own model evaluation.
"Feature Engineering and Selection: A Practical Approach for Predictive Models" by Max Kuhn and Kjell Johnson. This book covers practical techniques for feature engineering, which could help you refine your approach to improving model accuracy using SHAP or other methods.
"Machine Learning for Real Estate Valuation: A Review and Approaches" by Marzieh Babaeianjelodar et al. This review discusses different machine learning approaches used in real estate valuation, including feature engineering strategies specific to housing price prediction.
"Predicting House Prices Using Advanced Regression Techniques: A Kaggle Competition Approach". Although not a paper, the Kaggle competition discussions and top solutions often provide detailed insights into feature engineering and model selection strategies that work well for this dataset.
"Sustainability Assessment of Real Estate Developments Using Machine Learning Techniques" by Evrim Didem Gunes et al. This paper explores how machine learning can be applied to assess sustainability factors in real estate, which ties into ethical and sustainability considerations.

Professional Issues (Technical & Methodological Considerations)

Data Quality and Bias

- Housing data can have **incomplete records**, **measurement errors**, or **historical biases** (e.g., biased valuations in certain neighborhoods).
- Data sources (e.g., real estate platforms) may not always be **transparent**, raising reproducibility issues.

Model Transparency

- Predictive models, particularly **complex ones like gradient boosting or neural networks**, can be **black boxes**.
- Professionals are expected to document:
  - **Data sources**
  - **Feature selection methods**
  - **Model performance metrics (RMSE, MAE, etc.)**
  - **Model limitations**

Responsible Feature Engineering

- Ethical concerns arise if **sensitive features** (e.g., race, ethnicity, or proxies like ZIP code) are used directly or indirectly.
- Even non-sensitive features can become **proxies for race, socioeconomic status, or other sensitive attributes** — this is known as **proxy discrimination**.

2️⃣ Ethical Issues

Discrimination and Fairness

- AI systems for pricing could **reinforce existing biases** in housing markets.
- Example: If historical data reflects redlining (discriminatory practices in lending and pricing), the AI may **perpetuate those patterns**.

Transparency and Explainability

- Homebuyers, sellers, and policymakers may expect to understand **why a certain house gets a specific predicted price**.
- Ethical practice requires some level of **explainability**, especially in contexts like mortgage approvals.

Privacy

- Using personal data (e.g., past transactions linked to individuals) for modeling could raise **data protection** issues, particularly under laws like **GDPR**.

Accountability

- If your pricing model is used in **real estate platforms** or **valuation reports**, who is responsible for errors or biased predictions? Defining **accountability in AI-driven pricing** is an ongoing ethical debate.

3️⃣ Social Issues

Affordability and Housing Inequality

- AI models could be used to **justify inflated prices** in desirable areas, pricing out low-income families.
- There’s also the risk that AI-based valuations contribute to **gentrification** — displacing long-time residents.

Digital Divide

- Smaller, less technologically sophisticated property sellers (e.g., **older homeowners**) may be at a disadvantage if buyers rely heavily on AI predictions, reducing the **human negotiation element**.

4️⃣ Sustainability Issues

Environmental Impact of Data and Computing

- Training advanced models (e.g., ensembles, deep learning) can have a **large carbon footprint**.
- Sustainable AI encourages the use of:
  - **Simpler models when they suffice** (linear regression, decision trees).
  - **Efficient computing infrastructure**.
  - **Data minimization** (using only essential features).

Urban Planning and Resource Management

- Predictive models can support **sustainable housing policies**, helping identify:
  - Areas with excessive speculation.
  - Regions suitable for **eco-friendly development**.
  - Neighborhoods where **affordable housing policies** are needed.

House Prices Advanced Regression Techniques: A Step-by-Step Guide

Predicting house prices accurately is crucial for various stakeholders in the real estate market. In this blog post, we will walk you through a comprehensive approach to building a regression model that can predict house prices using advanced techniques. We will cover data loading, visualization, preprocessing, model building, and prediction.

📂 Step 1: Load and Explore the Data

First, we need to import the necessary libraries and load the datasets. This includes splitting the data into training and testing sets, and separating the target variable (SalePrice) from the predictors.

📊 Step 2: Visualizations

1. Distribution of Sale Prices
   Understanding how the sale prices are distributed can help identify skewness and outliers. By plotting the distribution of sale prices, we can visualize how prices are spread out and detect any anomalies or patterns that might affect our model.
2. Correlation Heatmap
   A heatmap of correlations between numerical features and the sale price helps identify which variables are most related to the price. This can guide feature selection and engineering by highlighting the most influential factors on house prices.
3. Boxplot of Sale Prices by Overall Quality
   This visualization shows the relationship between the overall quality of a house and its sale price. By plotting sale prices against the overall quality, we can see how different levels of quality affect prices, which can be a valuable insight for our model.

🛠️ Step 3: Preprocess the Data

We need to handle missing values, encode categorical variables, and scale numerical features to prepare the data for modeling. This involves:

🧹 Imputing missing values to ensure no gaps in the data.

🔤 Encoding categorical variables to convert them into a numerical format that can be used by machine learning algorithms.

📏 Scaling numerical features to standardize the range of values, which can improve the performance of certain algorithms.

🤖 Step 4: Define the Model and Bundle Preprocessing and Modeling Code in a Pipeline

We use a RandomForestRegressor for this task. By bundling preprocessing and modeling steps into a pipeline, we streamline the process and make it easier to fit the model to the training data and apply it to the test data.

🧾 Step 5: Predict and Prepare Submission

Predict using the test dataset and prepare the dataframe for submission. After training the model, we generate predictions for the test set. These predictions are then converted from the logarithmic scale back to the original scale, and a submission file is created for evaluation.