Final Project Report: House Price Prediction with Bias and Fairness Considerations

---

1. Introduction

The goal of this project was to develop a machine learning model to predict house prices, with a particular focus on mitigating biases related to location, socio-economic status, and house condition. By building a fair pricing model, we aimed to address ethical concerns in real estate pricing and ensure that the model does not disproportionately undervalue or overvalue properties based on non-economic factors.

2. Problem Framing

Problem Statement: We set out to develop a house price prediction model that would be fair across different neighborhoods and property types. Our focus was on identifying and mitigating biases associated with location, socio-economic status, and house condition. The unique angle of this project lies in the fairness considerations, ensuring that the model provides ethical pricing predictions across diverse demographic and socio-economic groups.

Assumptions: Several assumptions were made at the outset:

- **Data Quality:** The data provided was assumed to be accurate and representative of the housing market in Ames.
- **Missing Data:** We assumed that missing data could be handled without systematically biasing the predictions. We used imputation techniques where appropriate.
- **Feature Representation:** We assumed that the features provided were sufficient to capture the key factors influencing house prices and that no critical variables were omitted.
- **Fairness Definition:** Fairness in this context means avoiding systematic overvaluation or undervaluation of properties based on non-economic factors, particularly location.
- **Model Generalization:** We assumed that the trained model would generalize well to unseen data from the same housing market.

---

3. Data Preparation and Exploratory Data Analysis

We used the **Ames Housing Dataset**, which contains various features related to the characteristics of residential homes in Ames, Iowa. This dataset includes details on house size, condition, location, and year of construction, among other factors.

Data Cleaning and Handling Missing Data:

- We identified and imputed missing values using median imputation for numerical features and mode imputation for categorical features. This ensured that the dataset was complete and ready for modeling.
- We used PyJanitor for cleaning and standardizing column names, ensuring consistency across the dataset.

Exploratory Data Analysis (EDA): EDA was conducted to gain insights into the data distribution, key relationships between features, and target variable behavior. Key findings from the EDA included:

- **SalePrice Distribution:** We observed a slightly right-skewed distribution for `SalePrice`, which we considered transforming (e.g., log transformation) but ultimately decided to keep in its original form for model transparency.
- **Correlations:** We identified strong correlations between `SalePrice` and features such as `OverallQual`, `GrLivArea`, and `GarageCars`. These were expected to be key drivers of house prices.
- **Categorical Analysis:** We analyzed key categorical variables like `Neighborhood` and `OverallQual`. `Neighborhood` was found to be a significant feature, indicating that location played an important role in determining house prices.

---

4. Feature Engineering

Feature engineering was crucial in improving the model's predictive performance. We created new features to capture potential interactions and non-linear relationships between variables.

- **Interaction Features:** We engineered features such as `OverallQual * GrLivArea` and `TotalBsmtSF * GarageCars` to capture interactions between important predictors.
- **Polynomial Features:** We squared features like `GrLivArea` and `OverallQual` to capture non-linear effects that could influence house prices.

- **Temporal Features:** We created features like `HouseAge` and `YearsSinceRemodel` to capture the effect of time on house values.
- **Binned Features:** Continuous variables like `GrLivArea` were binned into categories (e.g., small, medium, large) to identify specific trends.

These newly engineered features allowed the models to better capture the complexity of house pricing, improving the overall accuracy of the predictions.

---

5. Modeling and Evaluation

We tested several machine learning models for house price prediction, including:

1. **Linear Regression**
2. **Decision Tree Regressor**
3. **Random Forest Regressor**
4. **Gradient Boosting Regressor**

Model Performance: After training and evaluating the models, the **Gradient Boosting Regressor** emerged as the best-performing model, achieving the highest R^2 score and the lowest error metrics. Here are the performance metrics across all models:

Model	MAE	MSE	RMSE	R^2
Linear Regression	0.0965	0.0454	0.2131	0.7566
Decision Tree	0.1440	0.0401	0.2004	0.7849
Random Forest	0.1006	0.0221	0.1486	0.8817
Gradient Boosting	0.0959	0.0195	0.1397	0.8955

Hyperparameter Tuning: To improve the performance of the Gradient Boosting model, we conducted hyperparameter tuning using `RandomizedSearchCV`. The best parameters identified were:

- **n_estimators:** 200
- **max_depth:** 3
- **learning_rate:** 0.05
- **min_samples_split:** 10
- **min_samples_leaf:** 1

These parameters improved the accuracy of the model while controlling for overfitting.

---

6. Feature Importance Analysis

Feature importance analysis was conducted to identify the most critical features in predicting house prices. The top contributors were:

- **OverallQual:** By far the most important feature, contributing over 40% to the predictions.
- **GrLivArea:** The second most significant feature, highlighting the impact of living area on house prices.
- **TotalBsmtSF and GarageCars:** Moderate contributors that indicate the value of basement space and garage capacity.

These findings aligned with our expectations from the exploratory data analysis and underscored the importance of house condition, size, and location in determining property values.

---

7. Fairness and Bias Considerations

While the primary goal was achieved in building an accurate and predictive model, **we did not fully address fairness and bias mitigation**. Although we evaluated feature importance and trained the model on diverse features, we did not conduct formal fairness audits or apply bias mitigation techniques.

- **Neighborhood Bias:** The `Neighborhood` feature, which was highly predictive of house prices, might act as a proxy for socio-economic status. Without fairness-aware algorithms, there is a risk of the model overvaluing or undervaluing houses based on their location.
- **Future Work:** Future efforts could involve conducting fairness audits and using techniques such as reweighting or adversarial debiasing to ensure that predictions are fair across different socio-economic groups.

---

8. Conclusion

The Gradient Boosting model successfully predicted house prices with high accuracy. The model identified `OverallQual`, `GrLivArea`, and `Neighborhood` as key drivers of house pricing. However, while the predictive performance was strong, there remains a need for fairness auditing to ensure that the model does not propagate socio-economic biases.

This project demonstrates the effectiveness of machine learning in real estate pricing and sets the foundation for future work on ethical and fair predictive modeling.

---

9. Future Work

Moving forward, we recommend:

- **Fairness Auditing:** Implement fairness metrics to ensure that the model's predictions are equitable across neighborhoods and socio-economic groups.
- **Bias Mitigation Techniques:** Explore bias mitigation techniques to reduce the risk of overvaluation or undervaluation in disadvantaged areas.
- **Additional Data:** Integrate external data (e.g., economic indicators, demographic data) to capture a broader range of factors influencing house prices.