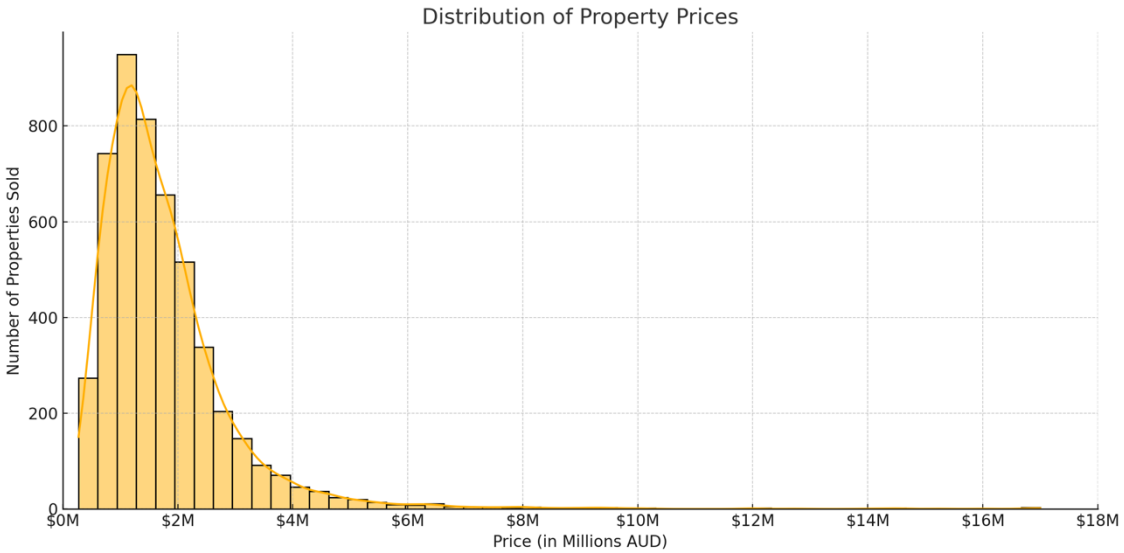
1. Challenges

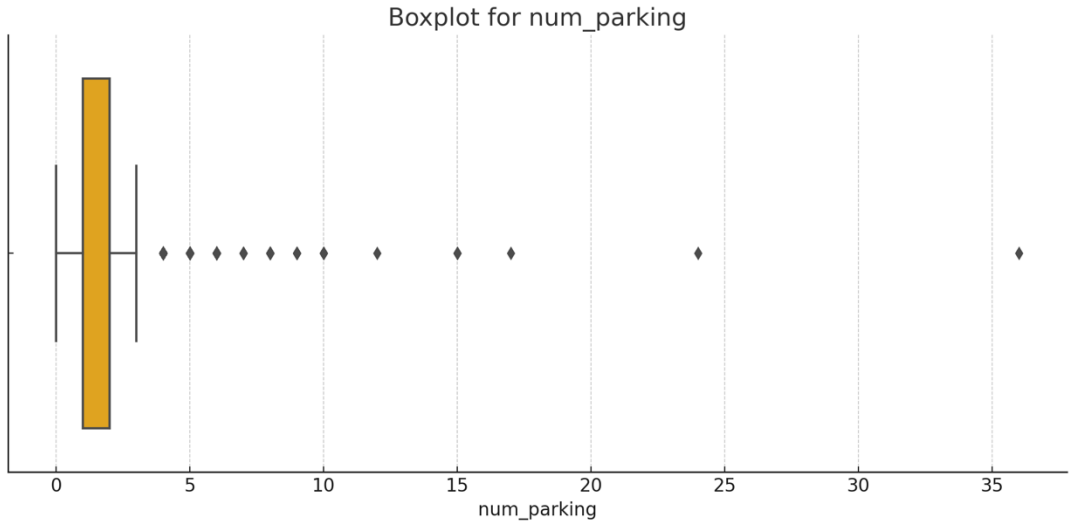
1.1 House Price Prediction (Regression)

Skewed Distribution

The housing price variable exhibits a strong right skew, with the majority of homes priced below $2M and a long tail stretching beyond $10M. This violates assumptions of models like linear regression and biases learning toward high-end outliers, making the model underperform for median homes.



There are also some outlier/noises in the data without considering the Price column. For example, there is a row of data that has even over 35 park spaces which is not common.

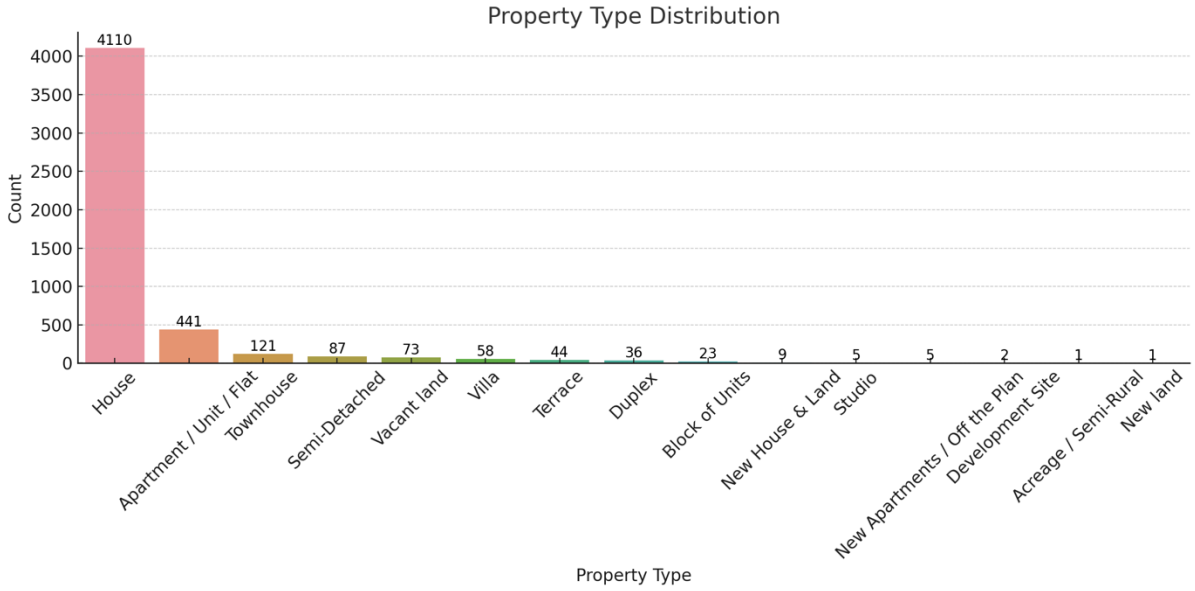


- **High Price Variance**: Prices range from $275K to $17M (σ = $1.25M) → large spread complicates learning.
- **Outliers**: Luxury properties skew model behavior.
- **Multicollinearity**: Many features (income, rent, suburb price) are correlated.
- **Sparse Regions**: Some suburbs and property types appear infrequently → harder for model to generalize.

1.2 Property Type Prediction (Classification)

Severe Class Imbalance

Over 80% of listings are "House", while other classes such as "Villa", "Studio", and "Acreage" appear infrequently. Models trained on this data tend to ignore rare classes unless weighted appropriately. A naive classifier predicting only “House” can achieve high accuracy yet fail completely on underrepresented categories.



Feature Space Overlap

Types like "Apartment" vs "Studio" or "Townhouse" vs "Villa" often share similar feature distributions (e.g., area, location). Without architectural metadata or zoning codes, the model struggles to distinguish structurally different yet statistically similar entries.

Label Noise

Some categories (e.g., “New House & Land”, “Block of Units”) reflect marketing jargon rather than objective structure, introducing inconsistencies in the target variable that complicate model convergence. Similar properties may be labeled differently (e.g., Duplex vs House).

1.3 Strategies to Address These Issues

Feature Engineering

- Interaction terms (distance × driving time, month_sin, day_cos) introduced to surface non-linear patterns.
- Applied log transformation to price for more stable regression learning.

Modeling Techniques

- Tree-based models (XGBoost and LightGBM) selected for their capacity to:
    - Handle high-cardinality categorical encodings,
    - Learn non-linear feature interactions,
    - Adapt to imbalanced classification via built-in class weighting.
    - I also tried random forest model but showed a lower F1 score compared to the current models, therefore I just ensembled the top two models.
    - In the final model,I ensembled the XGBoost and LightBGM models and achieved a higher F1 score while balancing the biases. (mentioned in Evaluation section)

Data Handling

- Dropped extremely rare target classes (frequency < 2) to avoid overfitting noise.
- Balanced class weights for classification; added temporal and structural signal to regression.
- Used random forest model to fill the null values based on the similar data instead of using mean or median .This apporach would be better than removing the entire row of data and is more accurate than mean or median filling.

1.4 External Data Sources for Enhanced Prediction

Source	Data Type	How It Helps
NSW Open Data	Auction rates, approvals	Model seasonal demand and developer pressure
ABS Census	Income, demographics	Add socio-economic context to suburb or LGA
RBA Economic Indicators	Mortgage rates, inflation	Adjust for affordability trends over time
OpenStreetMap / Transport NSW	Real-time commute networks	Replace estimated travel times with real transport data
Domain/RealEstate APIs	School catchments, amenities, crime	Enrich feature set for lifestyle and neighborhood valuation
Satellite Imagery APIs	Land use, tree cover, water access	Classify property environment (urban, suburban, coastal, etc.)

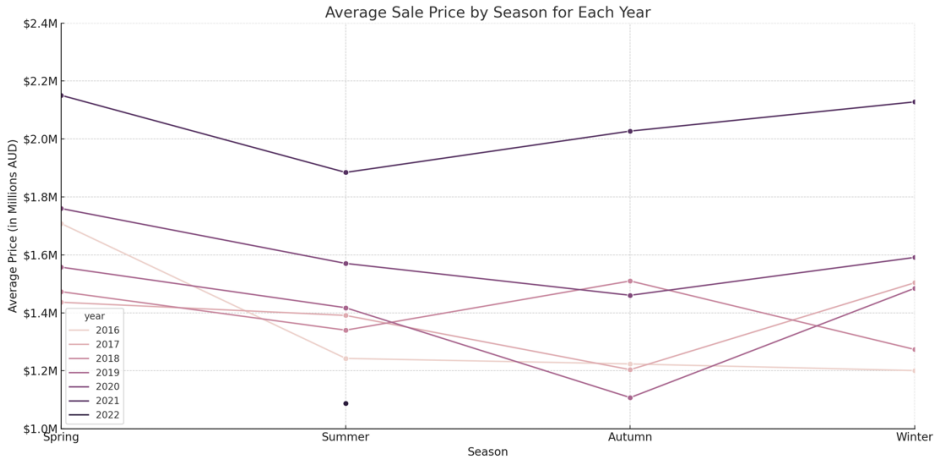
These sources would allow models to incorporate **external signals** that influence both market-level and property-specific decisions.

2. Incorporating Temporal Features

As mentioned above, real estate is a temporal market. Price, desirability, and availability shift with time, influenced by both macroeconomic and local trends.

2.1 Importance of Time Features

- Property markets are **seasonal** and **time-sensitive**.
- Prices surge in **Spring**,and often dip during the **Summer (as seen in the chart).**
- **Economic changes over years** (e.g., interest rate shifts) affect market prices.



2.2 Challenges in Encoding Time

Dates are **ordinal but not linear**,the difference between December and January is 1 month, not 11. Raw numerical month or day values introduce artificial boundaries that confuse models.

2.3 Solution: Temporal Feature Transformation

Feature	Role
year, month, day	Captures long- and short-term changes
quarter	Captures fiscal behavior or auction waves
month_sin, month_cos	Cyclic encoding to smooth temporal transitions
day_sin, day_cos	Captures intra-month fluctuations

These engineered features allow the model to learn **cyclical patterns** without discontinuity, supporting generalization across years.

3. Evaluation Metrics and Justification

3.1 Regression Metrics (House Price)

Metric	Description	Why Not Chosen?
MAE (✓ used)	Avg absolute error in dollars	Most interpretable, least sensitive to skew
RMSE	Penalizes large errors more	Biased by price outliers
R^2	Variance explained by model	Misleading under skewed distributions
Median Abs Error	Median error, more robust than MAE	Less intuitive to interpret across samples

MAE was chosen because price errors are best understood in real dollar terms, and it avoids the disproportionate influence of high-end outliers on model evaluation.

3.2 Classification Metrics (Property Type)

Metric	Description	Why Not Chosen?
Accuracy	% of correct predictions	Inflated due to majority class dominance
Precision / Recall	Class-specific performance	Helpful but incomplete alone
Macro F1	Unweighted F1 across all classes	Overweights rare classes
Weighted F1 (✓ used)	Weighted by support (class count)	Balances fairness and real-world relevance

Weighted F1 was selected because it ensures the performance on rare property types is not lost in aggregate accuracy, yet it doesn't exaggerate their importance.

3.3 Final Evaluation Scores

Task	Model	Internal Score	External Score
Classification (Final model)	Ensemble	F1 = 0.91	F1 = 0.906
Regression (Final model)	Ensemble	MAE ≈ $295k	MAE = $298,546

I also tired random forest model but did not work as good as LightGBM and XGboost models and when I removed it from the ensemble model the performance improved.

The random forest model achieved about 82% F1score and 330k MAE which were worst higher than the other two models (LightGBM -> F1:84.36 and MAE:301305 , XGBoost -> F1:89.5 and MAE:308599) .As a result,I removed it.