Technical Summary - Winnipeg Property Valuation Model

Preamble

The automated property valuation model (AVM) was built to predict assessed values for Winnipeg residential properties. In accordance to the assessment guidelines, the model build emphasized reproducibility, explainability, and operational readiness. The complete codebase, environment specifications, and configuration parameters are version-controlled on GitHub, with model development executed in Google Colab.
The python code implementation follows the R template structure provided by the examiner.

Data Acquisition and Preparation

Property assessment data was acquired via Socrata API from Winnipeg's Open Data Portal. The initial dataset contained 45 variables including the target (total_assessed_value) and various property characteristics. Data quality assessment revealed minimal missing data in the target variable (0.1%) but moderate missingness in features (1.8-16.3%). No duplicate records were identified.

Data preprocessing involved:

- Removing unhashable nested dictionary columns
- Dropping unique identifiers (gisid, roll_number, geographic coordinates)
- Converting object-type strings to numeric formats
- Identifying and removing data leakage (assess_value_1 was essentially the target)
- Multicollinearity assessment via Variance Inflation Factor analysis

Feature Engineering and Selection

Correlation analysis and domain knowledge guided selection of six numeric features:

- total_living_area (strongest predictor)
- year_built (property age indicator)
- rooms (capacity measure)
- assessed_land_area (land value component)
- water_frontage_measurement and sewer_frontage_measurement (property boundaries)

VIF analysis confirmed acceptable multicollinearity levels (VIF=9.83 for frontage measurements, below the 10.0 threshold). Features like street_number (identifier) and dwelling_unit (unavailable post-cleaning) were excluded.

Modeling Methodology

A 70/30 train-test split with random_state=42 ensured reproducibility. Three algorithms were evaluated:

1. **Linear Regression (Baseline):** R^2=0.83
   Provided interpretable coefficients and strong baseline performance, validating linear relationships between features and target.

2. **Random Forest (Selected):** R^2=0.86
   Ensemble of 100 trees (max_depth=20, min_samples_split=5) captured non-linear relationships and feature interactions. Outperformed alternatives through superior handling of complex patterns.

3. **XGBoost:** R^2=0.8377
   Gradient boosting with 100 estimators showed solid performance but marginally underperformed Random Forest, likely due to dataset size and feature complexity.

Model Selection Rationale

Random Forest was selected based on:

- Highest R^2 score (86% variance explained)
- Robustness to outliers and missing data
- Ability to capture non-linear relationships
- Built-in feature importance rankings
- No assumptions about data distributions

Key Findings and Insights

Feature importance analysis revealed total_living_area as the dominant predictor, followed by assessed_land_area and year_built. The most recent properties date to 2023, indicating current data. However, the model lacks renovation/expansion information, which could affect valuations when living area increases post-construction.

Geographic variables would likely improve predictions significantly, as real estate values are heavily location-dependent. Future work should incorporate one-hot encoded neighborhood features and proximity metrics (distance to downtown, schools, amenities).

Limitations

The 1,000-property sample may not capture full market diversity. Linear regression assumptions (homoscedasticity, normal residuals) were examined; Random Forest's non-parametric nature reduces sensitivity to violations.

Assumptions

Data dictionary of the 45 columns in the dataset was built on explanations/meanings provided by Chatgpt

Conclusion

The Random Forest model provides reliable automated valuations (R^2=0.86) suitable for deployment. Recommended enhancements include location features, expanded training data, and temporal analysis for market trend predictions.