# Technical Summary: Automated Property Valuation Model

## Assumptions

We assume that the Assessment Parcels data include high quality property level data (based on the quality rank). including numeric features (e.g., 'Rooms', 'Housing age/ Year Built', 'Sewer / Frontage Measure', 'Dwelling Units', 'Land ‘Living Area', 'Assessment Date', 'Centroid Lat' / Lon',) and categorical attributes (e.g., Building Type’, ‘AC, ‘Attached / Detached Garage’, ‘Pool', , ‘Basement and Finish’, ‘Fire Place', 'Multiple Residences’, ‘Market Region'). We also assume that the data exhibit a trend/pattern that can be learned to automate assessment on unseen data.

Numeric variables were standardized using StandardScaler to account for differences in scale and range across features. Categorical variables were encoded using Label Encoder, with missing values imputed using column-wise mode values. For experimentation, One Hot Encoding was also applied to evaluate its impact on model performance, particularly with linear regressors as they typically benefit from binary representation of categorical attributes. However, One Hot Encoding resulted in sparsity and degraded performance in tree-based models due to the high dimensionality of certain variables (e.g., Street Type – 38 categories, Street Name – 4,231, Property Use Code – 98, Neighborhood Area – 237). To protect privacy, all unique identifiers (e.g., property roll numbers) were removed to prevent data re-identification and linkage risks.

## Approach to the Business and Modeling Problem

The goal was to develop an Automated Valuation Model capable of predicting the assessed value of residential properties based on their physical, locational, and categorical attributes. This model will support municipalities and analysts in identifying under / over-valued properties, improving fairness and explainability in assessment. The workflow was designed for interpretability, justification, scalability, and reproducibility, allowing continuous re-training as new property data become available. Model performance was compared against the baseline linear model (R code) implementation provided, using R square as the primary evaluation metric.

## Summary of Applied Methods

### 1. Data Preprocessing:

Numeric and categorical columns were detected using a ETL pipeline. Numeric variables were scaled, and missing values were imputed using medians to reduce outlier effect (Very expensive / very low valued property). Outliers and records with illogical target values (e.g., assessed value of 0) were removed to improve data quality and R2 scores.

### 2. Model Training:

A variety of algorithms were implemented to identify the most suitable estimator:

- **Linear Models:** Ridge, Lasso, and Elastic Net were used to manage multicollinearity, perform automatic feature selection (Lasso), and combine L1/L2 regularization (ElasticNet).
- **Tree-Based Models:** Decision Trees were trained to enhance interpretability, followed by Random Forests to reduce overfitting via random bootstrapping and random feature selection.

- **Boosting Algorithms:** XGBoost was applied as a state-of-the-art gradient boosting model capable of handling non-linear relationships and identifying key predictors efficiently.
- Hyperparameter tuning was conducted using Grid Search and Randomized Search. Random search proved more efficient in exploring parameter space than exhaustive grid search.

**3. Model Evaluation and Visualization:**

Performance was evaluated using cross validation, R square and Root Mean Squared Error to measure explained variance and error. Feature importances, scatter/box plot and decision trees were visualized to interpret correlation and key drivers of assessed property values. One of the strengths of this approach is that **all the algorithms are tailored for the Winnipeg property evaluation** data so such well tailored algorithms can be utilized the evaluate / predict property value of Winnipeg in future.

**Model Comparison and Results**

Each model used consistent preprocessing and pre-split datasets for reproducibility and computational efficiency. The modular framework allows future retraining to mitigate data drift. Empirical testing showed:

- **Linear Models:** regularized linear model $R^2$ = 0.784 (356% improvement over the provided R baseline of 0.0034).
- **Decision Tree:** R2 improved by 2,500 % substantially but tended to overfit beyond depth 6.
- **MLP : R2** improved by 2,700 % substantially due to multiple layer based deep learning which can further be improved with larger data set and architecture search.
- **Random Forest:** Achieved $R^2$ = 0.934 with stable RMSE reduction (2,948% $R^2$ improvement).
- **XGBoost:** Delivered the highest $R^2$ (0.934) with the lowest RMSE, effectively capturing complex non-linear interactions.

**Conclusion**

The developed model provides an interpretable, high-performing, explainable and reproducible machine learning framework for large-scale property valuation. The use of scalable preprocessing, automated model comparison, and explainable tree-based methods ensures a transparent, fair, and data-driven assessment process suitable for municipal valuation systems.

***Notes***

*I used Google Collab for notebook-based coding and leveraged generative AI tool- Gemini to assist with code auto-completion and syntax suggestions, ensuring all implementations were written and customized by me.*