Vancouver Property Analytics: A Data Science Approach

Neel Sadafule, Saketh Poori, Sri Yenupothula

December 6, 2024

1 Introduction

Property tax levies and land values in Vancouver are important because they show market trends and impact investments, city planning, and public funds. Understanding what affects these values helps create fair taxes, better urban planning, and smarter decisions. This project uses statistics and machine learning to study and predict property taxes and house prices in Vancouver. By working with a dataset and using Apache Spark for fast processing, the project uncovers key patterns and insights for better predictions.

2 Methodology

This study is structured around three core components aimed at analyzing and predicting property tax levies and land values in Vancouver:

- **Data Preprocessing**: Cleaning and preparing the dataset to ensure consistency and reliability for analysis.
- **Feature Importance and Statistical Analysis**: Identifying significant predictors and exploring their relationships with property values and tax levies.
- **Predictive Modeling**: Developing and evaluating models, including Linear Regression, Polynomial Regression, Random Forest Regressor, and K-Nearest Neighbors Classifier, to forecast property tax levies and land values.

The objective is to uncover actionable insights into the factors driving tax assessments and property valuations, providing a deeper understanding of Vancouver's economic and urban landscape.

3 Data Preprocessing

Data preprocessing was essential to ensure the dataset's quality, consistency, and suitability for analysis and modeling. The original dataset contained **1,097,616** records, which was reduced to **693,881** after cleaning. To efficiently handle the large dataset, **Apache Spark** was utilized, leveraging parallel processing and all available CPU cores for fast data transformation.

Key preprocessing steps included:

- **Handling Missing and Duplicate Values**: Records with missing values in critical columns (`CURRENT_LAND_VALUE`, `PREVIOUS_LAND_VALUE`, `TAX_LEVY`) and duplicates were removed, significantly reducing the dataset size.
- **Outlier Treatment**: Outliers in financial metrics, such as `CURRENT_LAND_VALUE`, were capped using the Interquartile Range (IQR) method to prevent distortion in analysis.
- **Categorical Encoding**: Variables like `ZONING_DISTRICT` and `NEIGHBOURHOOD_CODE` were numerically encoded to enable inclusion in machine learning models.
- **Scaling and Log Transformation**: Logarithmic transformations were applied to skewed financial metrics (`CURRENT_LAND_VALUE`, `PREVIOUS_LAND_VALUE`) to stabilize variance. Numerical features were standardized to improve model performance.
- **Feature Engineering**: Derived features added new dimensions to the analysis:
  - `property_age`: Calculated from `YEAR_BUILT`, providing insights into the influence of property age on valuation and taxes.
  - `improvement_gap`: Representing changes in `CURRENT_IMPROVEMENT_VALUE` relative to `PREVIOUS_IMPROVEMENT_VALUE`, capturing property upgrades over time.

These steps ensured the dataset was clean, consistent, and optimized for reliable analysis and robust predictive modeling.

4 Feature Importance Analysis

Identifying factors influencing property tax levies and land value was the key objective of this study. This section discusses the methods used to find them.

4.1 Tax Levy

A Random Forest regression model was employed for its ability to handle non-linear relationships and provide interpretable insights into feature importance. The model combined numerical and categorical features into a single vector and was trained using `TAX_LEVY` as the target variable. Feature importance scores were then extracted and visualized to identify the most significant predictors.

- **Top Predictors**: `ZONING_DISTRICT` and `CURRENT_LAND_VALUE` emerged as the most critical factors influencing property tax levies. This aligns with domain knowledge, as zoning regulations and land value are pivotal in tax assessments.
- **Lesser Impact**: Features like `LEGAL_TYPE` and `improvement_gap` exhibited negligible importance, suggesting they could be deprioritized in future modeling efforts.

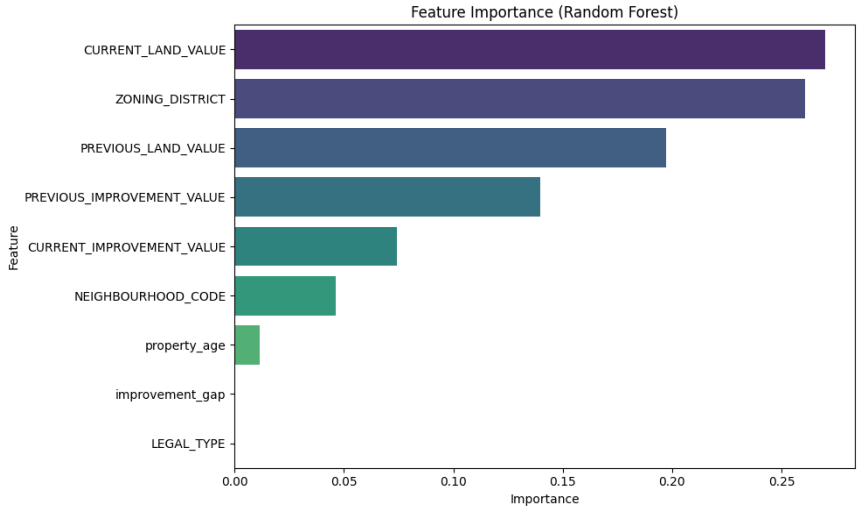

Figure 1: Feature Importance Analysis for Tax Levy Prediction (Random Forest).

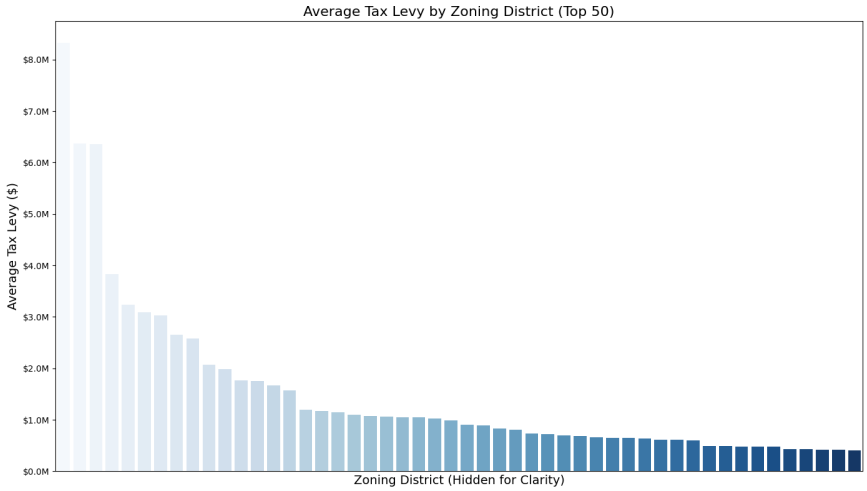

Figure 2: Average Tax Levy by Top 50 Zoning Districts.

Zoning and Tax Levy: Figure 2 highlights the top 50 zoning districts by average `TAX_LEVY`. These results underline the significant role zoning plays in determining tax levies, with certain zones contributing disproportionately to tax revenue.

4.2 Land Value

This section describes the important features for predicting land values.

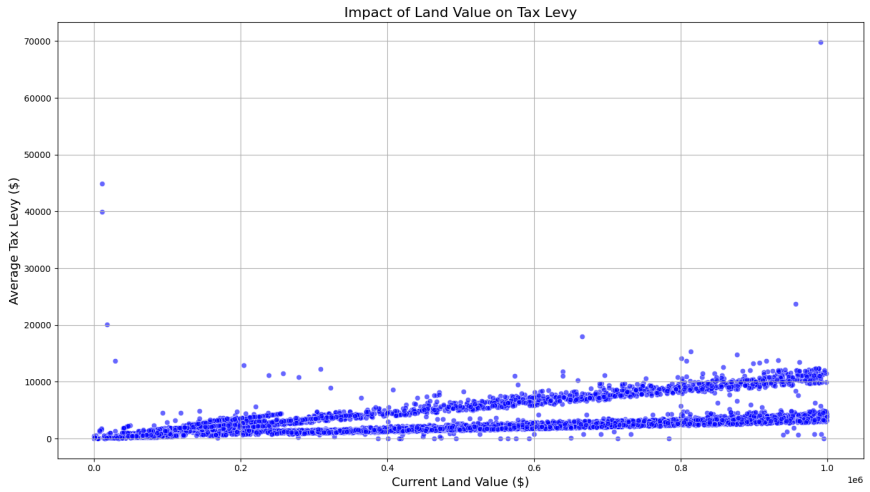

Figure 3: Scatterplot of Current Land Value vs. Tax Levy.

Figure 3 shows a positive correlation between `CURRENT_LAND_VALUE` and `TAX_LEVY`. However, properties with similar land values exhibit variability in tax levies, indicating that additional factors, such as zoning and neighborhood characteristics, also influence tax assessments.

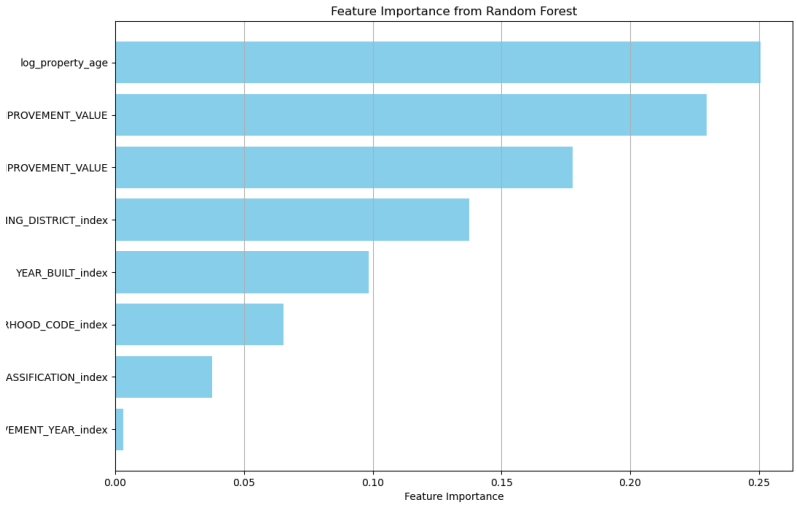

Figure 4: Feature Importance from Random Forest Regressor

Figure 4 displays the most influential features identified by the Random Forest Regressor. Notably, log property age and `CURRENT_IMPROVEMENT_VALUE` emerged as top predictors, although no single predictor dominated significantly.

4.3 Results

This analysis underscores the importance of zoning regulations and land value in property tax assessments. While these features dominate in their influence, the observed variability in tax levies for properties with similar land values points to the interplay of additional factors, such as neighborhood characteristics, that merit further exploration.

5 Statistical Analysis

To explore the relationships between `PREVIOUS_LAND_VALUE` and `NEIGHBOURHOOD_CODE`, statistical tests such as ANOVA and Tukey HSD were conducted. These tests aimed to determine whether significant differences exist in mean land values across different neighborhoods.

5.1 ANOVA Test

A one-way ANOVA test was performed to assess if there are significant differences in the mean `PREVIOUS_LAND_VALUE` among various `NEIGHBOURHOOD_CODE` groups. We filtered out rows with null values in the `PREVIOUS_LAND_VALUE` column. Then, we grouped the data by `NEIGHBOURHOOD_CODE` and collected land values for each group. Finally, we applied ANOVA to evaluate differences in group means. The ANOVA test resulted in a p-value of 0.0, which is significantly below the 0.05 threshold, indicating substantial differences in mean land values across neighborhoods.

5.2 Tukey HSD Test

Following the ANOVA test, a Tukey HSD (Honestly Significant Difference) test was conducted to identify specific neighborhood pairs with significant differences in mean land values. We created a contingency table with pairs of `NEIGHBOURHOOD_CODE` and their corresponding `PREVIOUS_LAND_VALUE`. Then we performed the Tukey HSD test with a family-wise error rate (FWER) of 0.05. From that we were able to identify neighborhood pairs exhibiting significant differences in mean land values. Our results showed that out of 435 comparisons, there were significant differences in 320 pairs and no significant differences in 115 pairs. These results confirm that neighborhood plays a crucial role in determining `PREVIOUS_LAND_VALUE`, with numerous pairs showing significant differences.

5.3 Chi-Square Test Analysis

We had several chi-square tests that we used in order to analyze our data better. We ran 4 different tests:

- Test 1: evaluating the association between `ZONING_DISTRICT` and `PRICE_CATEGORY`. The results indicate a strong association, suggesting that zoning regulations significantly influence property price categories.
- Test 2: exploring the relationship between the year a property was built (`YEAR_BUILT`) and its price category (`PRICE_CATEGORY`). The findings highlight a significant relationship, with older properties influencing price categories differently.

- Test 3: assessing the association between categorized land values (`LAND_VALUE_CATEGORY`) and tax categories (`TAX_CATEGORY`). The analysis confirms a strong dependency of tax categories on land valuations.
- Test 4: examined the relationship between `ZONING_CLASSIFICATION` and `TAX_CATEGORY`. The findings emphasize the influence of zoning classifications on tax systems and their critical role in urban planning.

.	Test 1	Test 2	Test 3	Test 4
Chi-Square Statistic	504,972.07	220,256.12	391,553.71	89,753.28
P-value at 5% sig level	0.0	0.0	0.0	0.0
Degrees of Freedom	368	228	4	20

Table 1: Results of Chi-Square Tests for Various Relationships in the Dataset

5.4 Results

The statistical tests reveal that both neighborhood and zoning significantly influence land values and tax categories. While ANOVA and Tukey HSD highlighted differences in land values, Chi-Square tests underscored strong dependencies between zoning classifications, land values, and tax categories. These results reinforced the importance of zoning and neighborhood factors in property assessment and taxation.

6 Predictive Model using Machine Learning

6.1 Baseline Model: Linear Regression

A Linear Regression model was established as a baseline for predicting property tax levies due to its interpretability and straightforward implementation.

6.1.1 Procedure

1. **Log Transformation**: Applied to skewed numerical features to enhance model stability.
2. **Feature Scaling**: Standardized all features to ensure uniform influence during training.
3. **Regularization**: Employed L2 regularization (Ridge Regression) to mitigate overfitting and improve generalization.
4. **Train-Test Split**: Divided the dataset into 80% training and 20% testing subsets to evaluate model performance.

6.1.2 Performance Metrics

- **Root Mean Square Error (RMSE)**: 1246.06
- **Mean Absolute Error (MAE)**: 439.36
- **R^2 (Coefficient of Determination)**: 0.62

The baseline model explained approximately 62% of the variance in property tax levies, providing a solid foundation for further model enhancements.

6.1.3 Visualizations

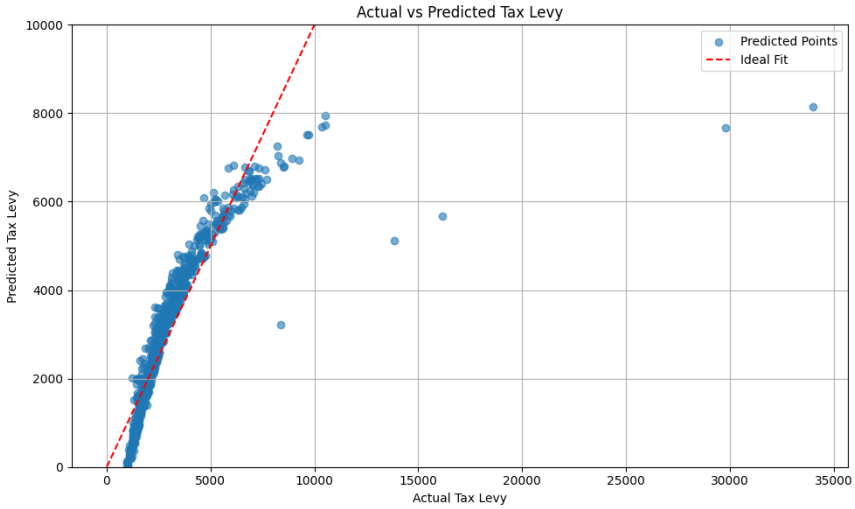

Figure 5: Actual vs Predicted Tax Levy (Linear Regression)

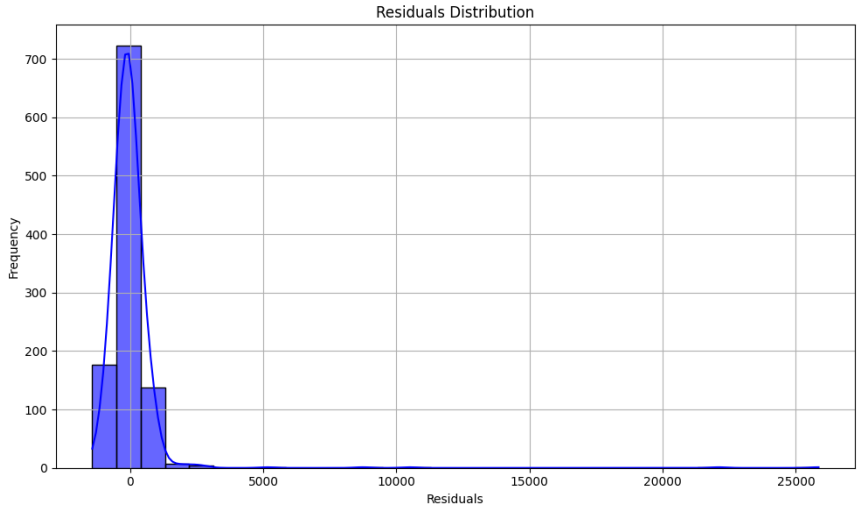

Figure 6: Residuals Distribution (Linear Regression)

Figure 5 compares actual tax levies with those predicted by the Linear Regression model, while Figure 6 displays the distribution of residuals, highlighting areas where the model's predictions deviate from actual values.

6.2 Improved Model: Polynomial Regression

To capture non-linear relationships and enhance predictive performance, a Polynomial Regression model was implemented by incorporating polynomial interactions.

6.2.1 Procedure

1. **Polynomial Expansion**: Generated feature interactions up to the second degree to capture non-linear relationships.
2. **Log Transformation**: Applied to skewed numerical features to stabilize variance.
3. **Feature Scaling**: Standardized polynomial features to ensure uniform influence during training.
4. **Regularization**: Utilized L2 regularization (Ridge Regression) to manage increased model complexity.
5. **Train-Test Split**: Maintained an 80% training and 20% testing split for evaluation.

6.2.2 Performance Metrics

- **Root Mean Square Error (RMSE)**: 1023.58
- **Mean Absolute Error (MAE)**: 331.38
- **R^2 (Coefficient of Determination)**: 0.74

The Polynomial Regression model improved the explained variance to 74%, showcasing its ability to capture complex data interactions more effectively than the baseline model.

6.2.3 Visualizations

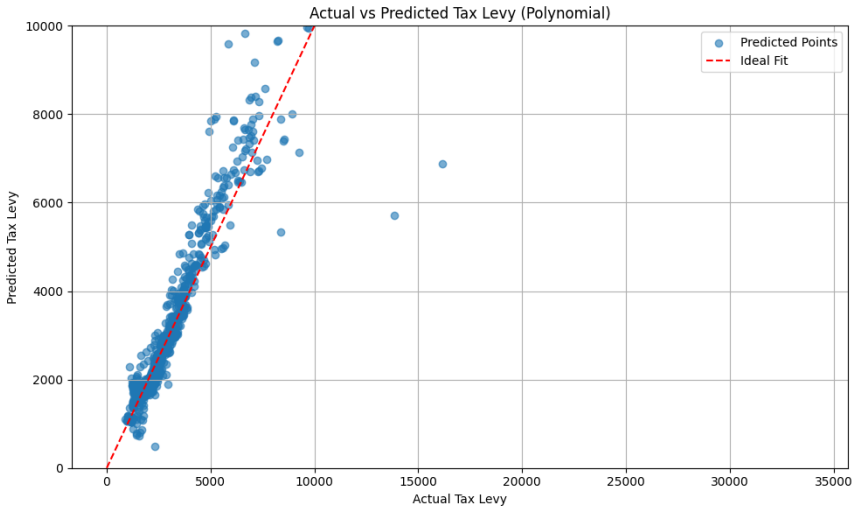

Figure 7: Actual vs Predicted Tax Levy (Polynomial Regression)

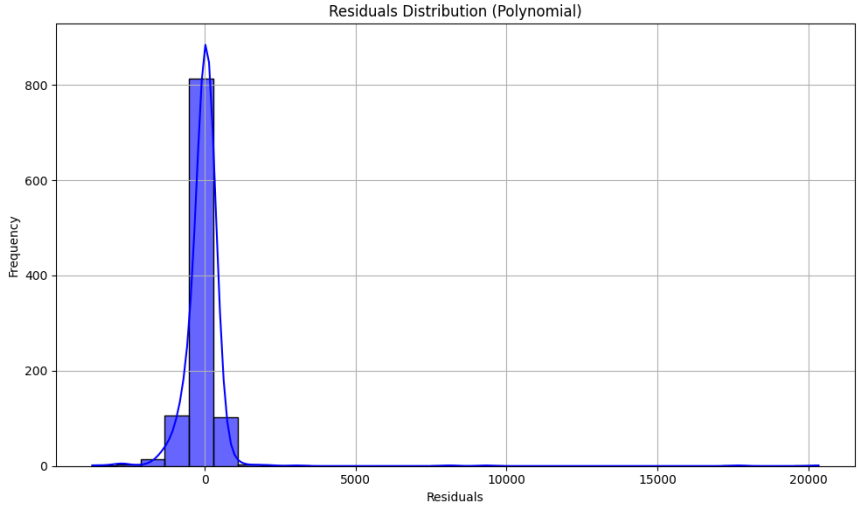

Figure 8: Residuals Distribution (Polynomial Regression)

Figure 7 illustrates the enhanced alignment between actual and predicted tax levies, while Figure 8 shows a more concentrated residual distribution around zero, indicating improved model accuracy.

6.2.4 Key Insights

The analysis revealed that incorporating polynomial features significantly enhanced the model's ability to capture complex interactions, leading to improved predictive performance. Critical drivers, such as `ZONING_DISTRICT` and `CURRENT_LAND_VALUE`, remained consistent as influential predictors

in the enhanced model. Additionally, the residuals demonstrated a better fit, with errors more closely centered around zero, indicating increased accuracy and reduced bias in the predictions.

6.3 Random Forest Regressor: Predicting Current Land Value

A Random Forest Regressor was utilized to predict `CURRENT_LAND_VALUE` of properties. This model was selected for its ability to handle large datasets, manage non-linear relationships, and evaluate feature importance effectively.

6.3.1 Procedure

1. **Data Preparation**:
    - Cleaned the dataset by addressing missing values and ensuring proper formatting of numeric features.
    - Hot-encoded categorical variables such as `ZONING_DISTRICT` and `NEIGHBOURHOOD_CODE` using `StringIndexer`.
    - Normalized numerical features and applied transformations where necessary, such as logarithmic transformation for property age.
2. **Feature Engineering**:
    - Engineered features like log-transformed property age and combined improvement values to capture more nuanced relationships.
3. **Model Training and Testing**:
    - Split the data into training (80%) and testing (20%) sets.
    - Trained a Random Forest Regressor on the training data.
    - Evaluated the model using key metrics: RMSE, MAE, and R^2.

6.3.2 Results

Figure 9 presents the residuals distribution, which appears approximately normal with most errors centered around zero. This suggests that the model is well-calibrated and effectively captures the underlying patterns in the data.

- **Root Mean Square Error (RMSE)**: 549,109.08
- **Mean Absolute Error (MAE)**: 388,591.46
- **R^2 (Coefficient of Determination)**: 0.6818

The Random Forest Regressor performed effectively in predicting `CURRENT_LAND_VALUE`, achieving an R^2 of 0.6818. The feature importance analysis highlighted key drivers of land value, providing valuable insights for stakeholders involved in property valuation and taxation. While the model shows strong performance, further refinements and the inclusion of additional data could enhance prediction accuracy.

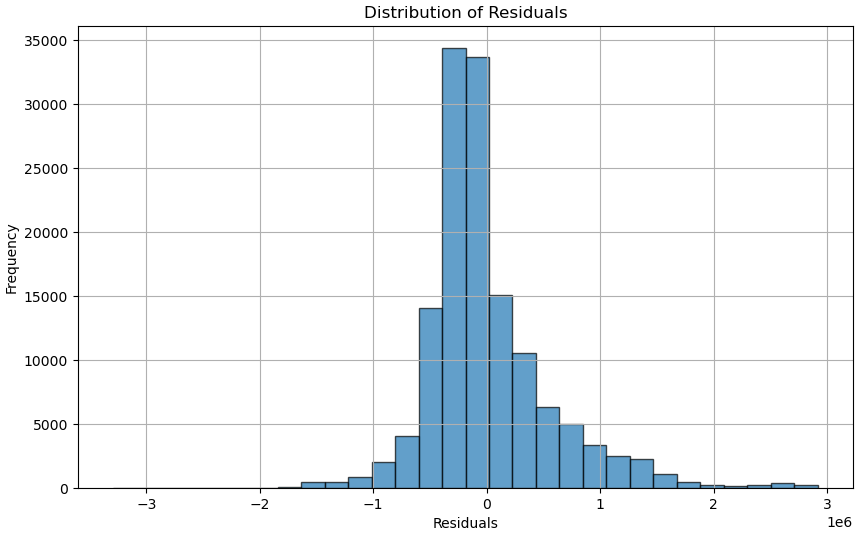

Figure 9: Distribution of Residuals

6.4 K-Nearest Neighbors (KNN) Classifier Analysis

A K-Nearest Neighbors (KNN) classifier was implemented to predict the `PRICE_CATEGORY` of properties based on a combination of numerical and categorical features. This analysis evaluates the model's performance and identifies challenges encountered during implementation.

6.4.1 Procedure

1. **Data Preparation**:
   - Removed duplicate records and cleaned column names by stripping leading/trailing whitespaces.
   - Converted key numerical columns (`CURRENT_LAND_VALUE`, `CURRENT_IMPROVEMENT_VALUE`, `PREVIOUS_IMPROVEMENT_VALUE`, and `property_age`) to double precision for consistency.
   - Encoded the target variable (`PRICE_CATEGORY`) using `StringIndexer` to create a `label` column.
   - Encoded categorical features (`ZONING_DISTRICT`, `NEIGHBOURHOOD_CODE`, `YEAR_BUILT`, `BIG_IMPROVEMENT_YEAR`, and `ZONING_CLASSIFICATION`) into numerical indices using `StringIndexer`.
2. **Feature Engineering**:
   - Combined numerical and encoded categorical features into a single `features_raw` vector using `VectorAssembler`.
   - Converted the Spark DataFrame into a Pandas DataFrame for further processing.
3. **Model Training and Evaluation**:

- Split the data into training (80%) and testing (20%) sets.
- Standardized the features using `StandardScaler`.
- Trained a KNN classifier with $k = 5$ on the scaled training data.
- Evaluated the model on the test set, calculating accuracy as the primary metric.

6.4.2 Results

The KNN classifier achieved an accuracy of 6.18% on the test data, indicating low predictive performance for this classification task. The low accuracy of the KNN classifier suggests that predicting `PRICE_CATEGORY` using the selected features and KNN methodology is challenging. Potential reasons include high dimensionality, irrelevant features, or insufficient data preprocessing.

7 Overall Conclusion

This report presented a comprehensive analysis of property tax and land value prediction, leveraging both statistical methods and machine learning models to derive key insights. Feature importance analysis revealed that land value, improvement value, and zoning were critical predictors of property values, while factors such as property age had a lesser impact, suggesting they may be deprioritized in future studies. The statistical analysis highlighted significant relationships between property features and values. Chi-Square and ANOVA tests revealed strong associations, particularly between zoning, neighborhood, and property values. Tukey HSD analysis further identified specific neighborhood pairs with notable differences, emphasizing the influence of local factors on property values. In predictive modeling, the Random Forest Regressor effectively captured complex relationships, achieving an R^2 of 0.68 for predicting `CURRENT_LAND_VALUE`. The Linear Regression baseline model provided a solid foundation at $R^2 = 0.62$, which was further improved to 0.74 using Polynomial Regression. However, the KNN Classifier demonstrated limited effectiveness, with an accuracy of only 6.18%, indicating the need for alternative approaches.

7.1 Limitations

We encountered several limitations when implementing this project. Data constraints posed a significant challenge, as access to comprehensive real estate data was limited, with some datasets locked behind paywalls or containing missing information. Computational challenges also emerged, with high memory and CPU requirements restricting the complexity and scale of models that could be employed. Furthermore, model limitations were evident, as certain models, like KNN, underperformed, highlighting potential inadequacies in feature selection or model suitability.

7.2 Future Work

Future studies could address these limitations and expand upon the current findings in several ways. Incorporating additional features such as economic indicators, environmental factors, and more granular geographical data could significantly enhance model accuracy. Exploring advanced models, like Gradient Boosted Trees, Deep Learning techniques, or ensemble methods, may further improve predictive performance. Expanding the data by including information from other regions or extending the temporal range could provide deeper insights.

7.3 Closing Remarks

This study establishes a foundational understanding of property taxation and valuation, offering valuable insights for urban planners, policymakers, and researchers. It identifies key factors influencing property values and tax levies while suggesting avenues for future research to further enhance predictive capabilities.

8 Project Experience Summary

8.1 Sri Hitesh Yenupothula, 301471359

Conducted statistical analyses including ANOVA, Tukey HSD, and Chi-Square tests using Python to identify key relationships between property attributes like neighborhood codes and zoning districts with property values, leading to the discovery of statistically significant differences across 320 out of 435 neighborhood pairs.

Developed and evaluated predictive models such as a Random Forest Regressor and KNN Classifier using Apache Spark and Pandas, achieving an R^2 of 0.6818 with Random Forest for land value prediction and identifying improvement value and zoning district as the most influential features.

Applied feature importance analysis using Random Forest to visualize and quantify the impact of predictors, offering actionable insights for stakeholders and improving model interpretability for practical applications in property valuation.

Co-authored the report summarizing statistical analyses, predictive modelling, and key findings, including visualizations and insights derived from Random Forest and KNN models, to effectively communicate project outcomes to both technical and non-technical stakeholders.

8.2 Saketh Poori, 301575678

Created predictive models for tax levy analysis using Apache Spark, including Random Forest and Linear Regression models with advanced preprocessing techniques such as log transformation, feature scaling, and L2 regularization, achieving an R^2 of 0.62 for tax levy predictions and extracting actionable insights on key drivers like land value and zoning.

Developed a data pipeline for feature importance extraction and tax levy prediction using PySpark, integrating data preprocessing steps like handling missing values, encoding categorical variables, and splitting datasets for efficient model training and evaluation.

Generated visualizations such as feature importance bar plots and residual analysis charts using Matplotlib and Seaborn, enabling stakeholders to interpret predictive model results and identify areas for further improvement in property valuation.

Co-authored the report summarizing statistical analyses, predictive modelling, and key findings, including visualizations and insights derived from Linear Regression models, to effectively communicate project outcomes to both technical and non-technical stakeholders.

8.3 Neel, 301418496

Preprocessed and cleaned large datasets by handling missing values, encoding categorical variables, and applying log transformations to stabilize skewed data distributions, ensuring the datasets were suitable for advanced machine learning algorithms and statistical testing.

Implemented feature engineering techniques to create derived variables such as property age and improvement gaps, capturing additional relationships and improving model interpretability in property tax and value predictions.

Conducted Chi-Square tests to evaluate the relationships between zoning districts, construction year, and property price categories, uncovering strong associations with p-values below 0.05 and emphasizing the role of zoning regulations and construction year in shaping property trends.

Improved linear regression models through iterative refinements, reducing RMSE from 1246.06 to 1023.58 and MAE from 439.37 to 331.38, while increasing R^2 from 0.62 to 0.74 increasing the reliability of the model.

Coordinated visualization and reporting efforts by integrating findings from statistical and machine learning analyses into a comprehensive technical report, effectively communicating results and actionable insights to technical and non-technical audiences.

Co-authored a comprehensive technical report summarizing statistical analyses, predictive modelling, and key findings, including visualizations and insights derived from Random Forest and KNN models, to effectively communicate project outcomes to both technical and non-technical stakeholders.

9 References

- Data Acquired from City of Vancouver Open Data Portal, *Property Tax Report*, `https://opendata.vancouver.ca/explore/dataset/property-tax-report/table/?sort=-tax_assessment_year`