Can machine learning help cities cut through the red tape of building permits?

Classifying Permit Types from San Francisco Building Permit Data Using Machine Learning

Benoit Loze

18/07/2025

Problem Statement

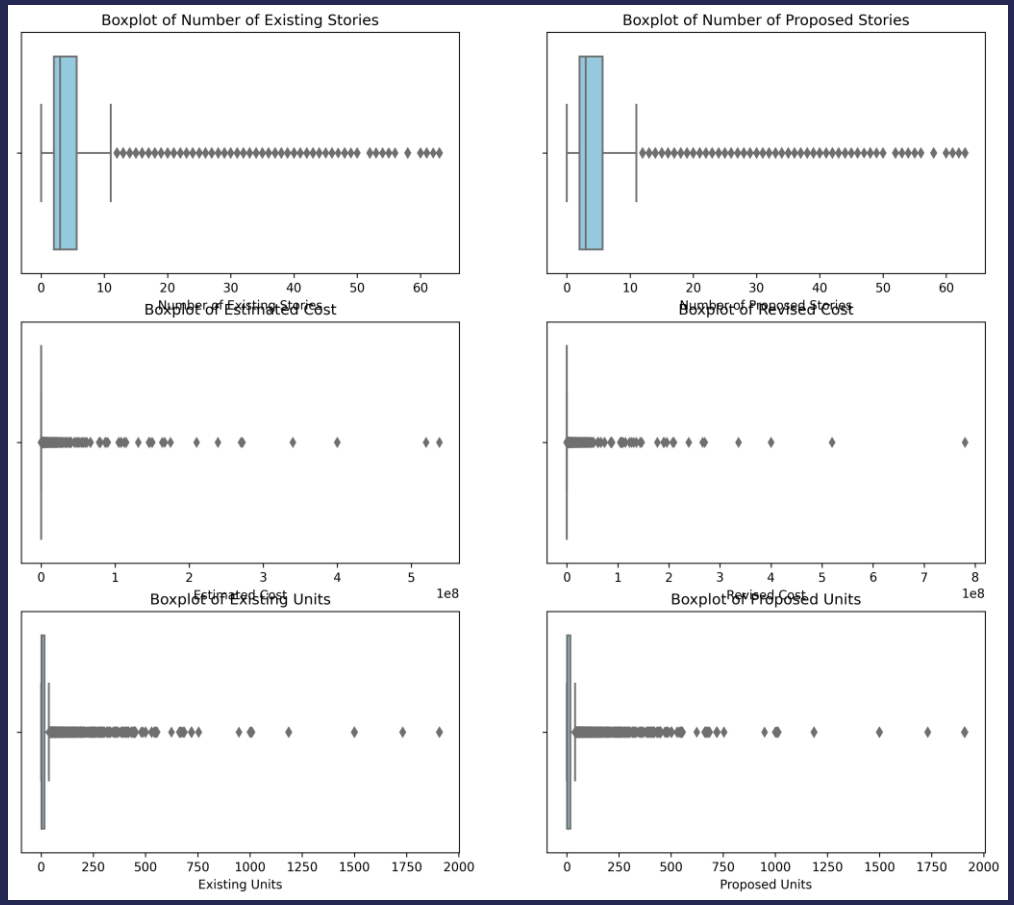
- **Goal:** Classify SF building permit types using structured features.
- **Success Criteria:** ≥80% accuracy.
- **Scope:** Supervised classification.
- **Constraints:** Missing data, class imbalance, feature engineering.

Data Wrangling

- Reviewed dataset (2013–2018), removed high-null columns.
- Imputed missing values.
- Converted date types and fixed categorical features.
- Saved cleaned CSV for next steps.

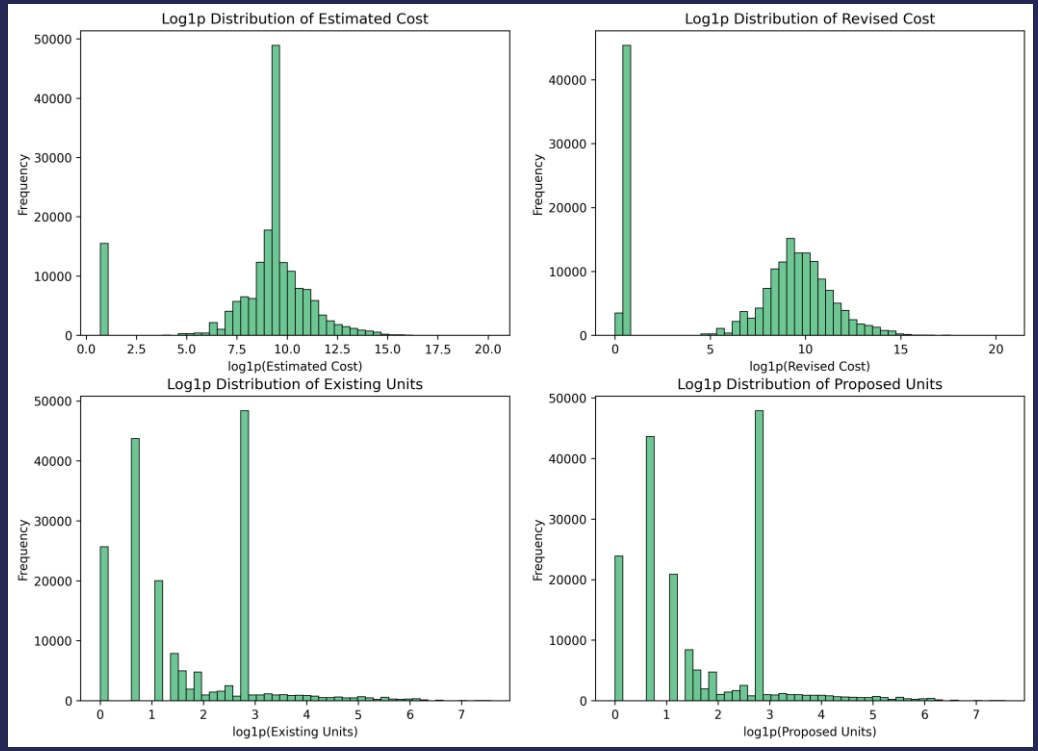
Outliers and Redundancies

- Identified outliers in numeric features.
- Visualized with boxplots and histograms.
- Removed redundant features and logged partial duplicates.



EDA Overview

- Normalized skewed variables with log1p.
- Split data by high-rise vs low-rise.
- Feature significance tested ($p<0.05$).
- Dropped collinear features to improve model reliability.

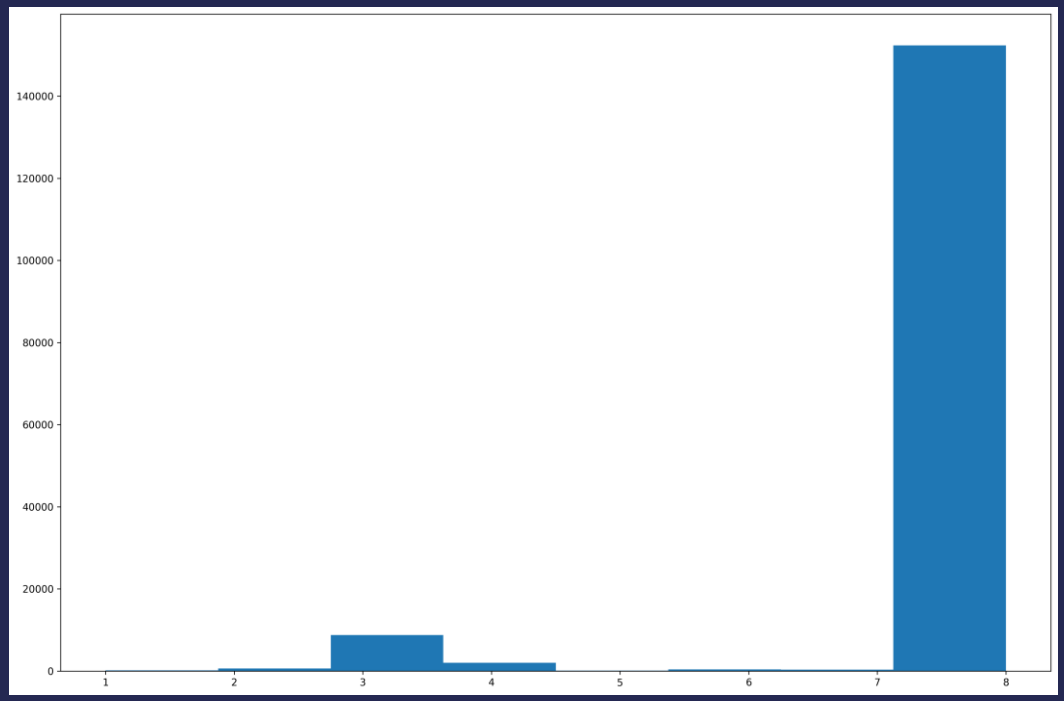


Categorical Analysis

- Explored distributions of categorical variables.
- Saved for next stages.

Data Preprocessing

- Handled class imbalance
- Encoded features per model type.
- Scaled data for regression models.
- Ensured no data leakage throughout pipeline.



Target Imbalance

- Severe imbalance: Class 7 ≈ 90%.
- Stratified data split ensured class representation.
- Used class_weight for rebalancing during model training.

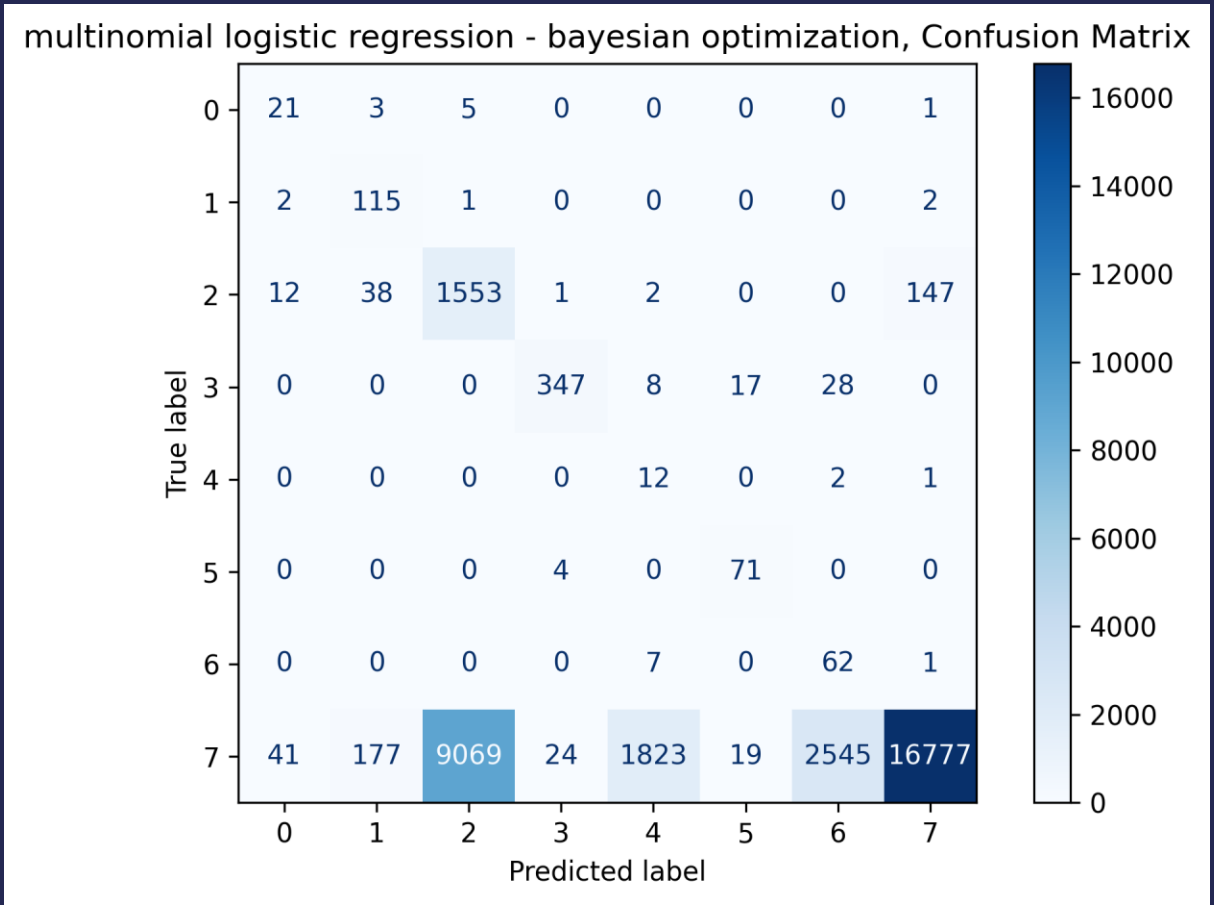
Modeling Approach

- Tested Logistic Regression and Random Forest.
- Evaluated via precision, recall, F1 score, confusion matrix.
- Avoided misleading accuracy metric.

Logistic Regression

- Used multinomial logistic regression.
- Bayesian optimization or Random search = same results.
- Still misclassified many samples due to class imbalance.

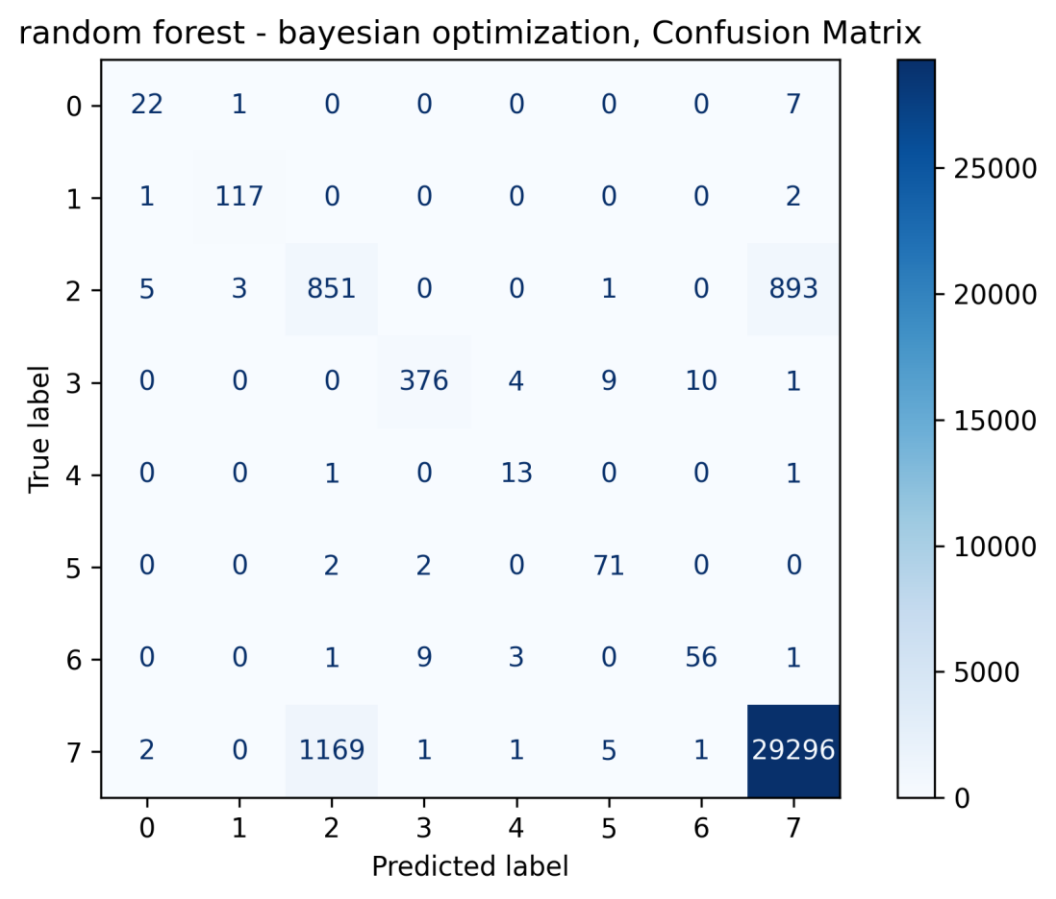
Logistic Regression Confusion Matrix



Random Forest

- Random Forest with Bayesian optimization yielded best performance.
- Better balanced precision and recall.
- Improved Class 7 prediction, fewer misclassifications.

Random Forest Confusion Matrix



Final Model Selection

- Random Forest (Bayesian Optimization)
- F1: 0.81, Precision: 0.79, Recall: 0.84
- Best handling of class imbalance.
- Robust across all permit types.

Conclusion

- Efficient pipeline designed to prevent leakage.
- Models tested and evaluated across real-world constraints.
- Outcome: Reliable automation of permit type classification using ML.