Exploratory Data Analysis of San Francisco Building Permit Data

Date Of Presentation	• 01-04-2025
Name	• Sunkara Harika
Pace Email Address	• hs21496n@pace.edu
Class Name	• Practical Data Science
Program Name	• MS in Data Science Seidenberg School Of Computer Science And Information Systems Pace University

Agenda

- Executive Summary
- Project plan recap
- Data
- Exploratory Data Analysis(EDA)
- Modeling methods
- Findings
- Business Recommendations and technical next steps

Executive summary

Problem:
The goal of this project was to analyze San Francisco's building permit system to uncover the factors contributing to delays and inefficiencies in the approval process. We aimed to build predictive models that could forecast approval outcomes and timelines, ultimately helping the city streamline operations, improve applicant experience, and allocate resources more effectively.

Solution:
We performed in-depth Exploratory Data Analysis (EDA) on over 190,000 permit records and applied machine learning models to:

- Predict whether a permit will be approved using project features
- Estimate the time required for approval
- Segment permits by complexity using clustering
- Visualize neighborhood-level trends and project patterns

This approach provides actionable insights to optimize city workflows, reduce processing time, and increase transparency in permit reviews.

Project plan recap

Deliverable	Due Date	Status
Data & EDA	03/25/2025	Completed
Methods, Findings, and Recommendations	04/01/2025	Completed
Final Presentation	04/22/2025	In Progress

Data

Data

- **Dataset:** San-Francisco Dataset
- **Source:** Kaggle
- **Sample Size:** 1,338 rows
- **Time Period:** March 28, 2012 to December 29, 2017. (Covers permit filings from early 2000s to recent years.)
- **Data that was purposefully Included or Excluded**:
    - The dataset includes permit type and status, filing, issue, and completion dates, estimated and revised project costs, project scope details such as number of stories and units, as well as neighborhood and construction type information.
    - Personally identifiable information (PII), contractor names (due to inconsistent availability), and sparse columns like TIDF Compliance and Soft-Story Retrofit were excluded during cleaning due to high rates of missing data.
- **Assumptions:**
    - We assumed that dates such as Issued Date and Filed Date are correctly and consistently formatted, the Estimated Cost accurately reflects the project scope, missing values in non-critical fields (like Unit Suffix or Site Permit) do not significantly impact model predictions, and the Final Status field reliably represents the actual outcome of the permit application.

Exploratory Data Analysis(EDA)

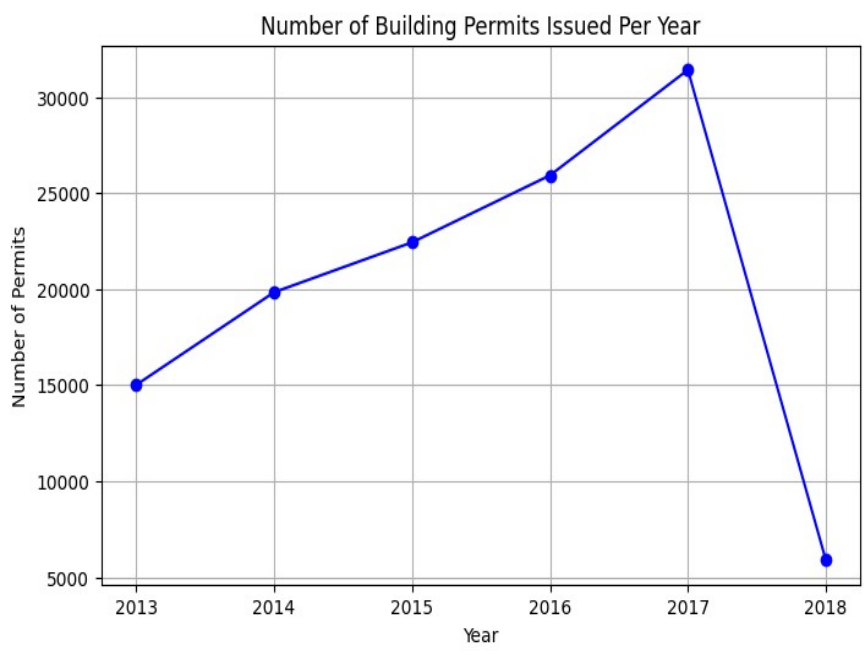
Permit Trends Over Time

Key Takeaways:

- Permit issuance steadily **increased from 2013 to 2017**, peaking at over 32,000 permits in 2017.
- This upward trend reflects a period of high construction activity and demand in the city.
- In 2018, there is a sharp drop, with permits falling below 10,000 — a possible indicator of data gaps, policy changes, or slowed development.
- The trend shows strong growth for five years followed by an unexpected drop in the sixth.

Data Notes:

- Source: Kaggle San Francisco Dataset
- Time Period: 03-28-2012 to 12-29-2017
- Purpose: Shows year-wise construction activity trend.
- Insight: Strong growth through 2017, sharp fall in 2018 (possible data gap or market shift).



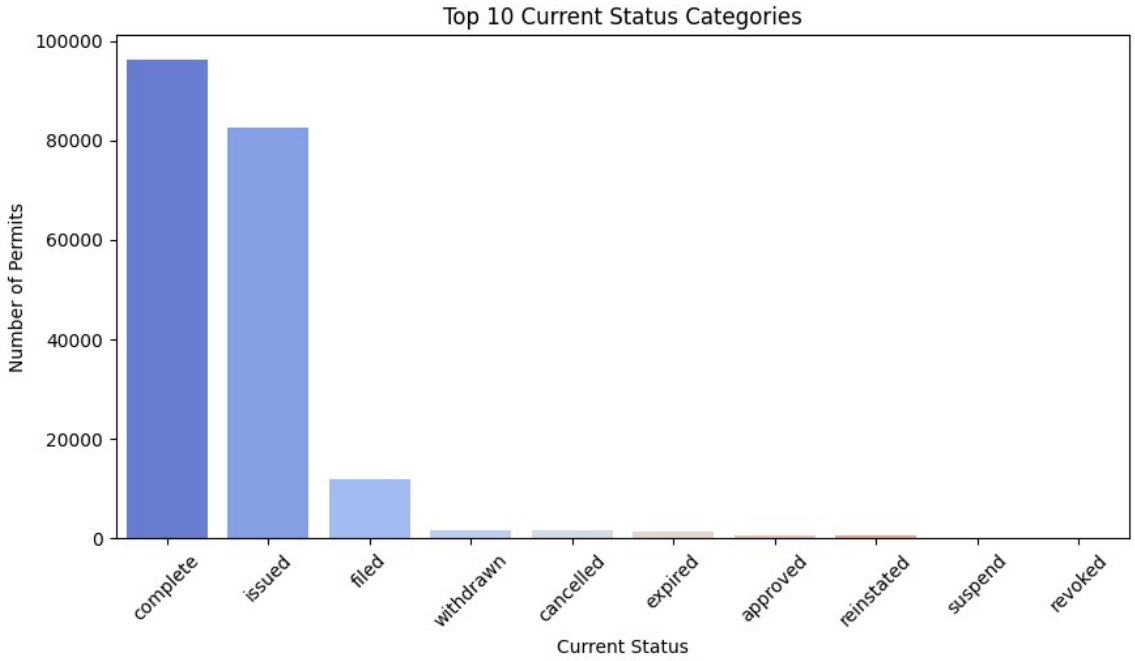
Permit Status Distribution

Key Takeaways:

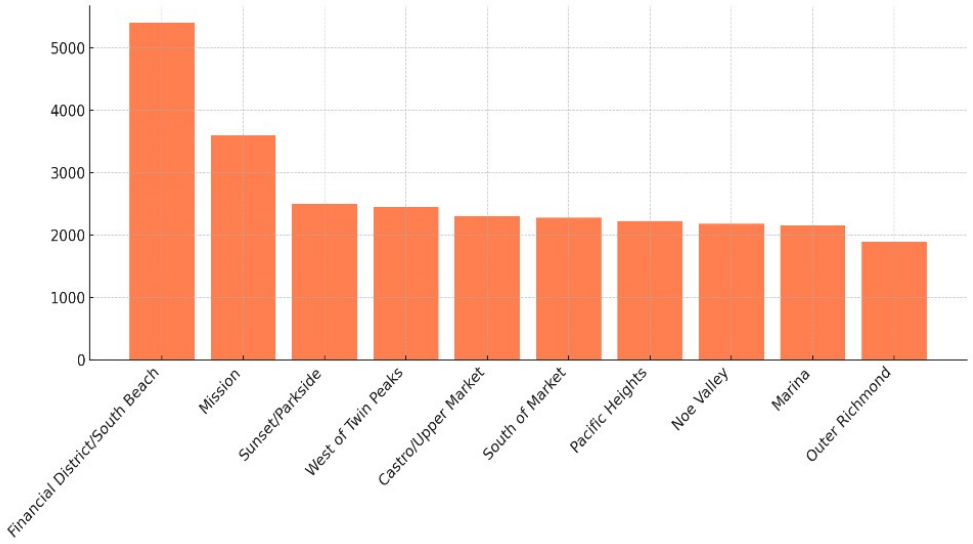
- The majority of permits are marked "complete" and "issued", indicating successful progression through the system.
- A smaller portion is still in the "filed" stage, possibly pending review or awaiting approval.
- Withdrawn, cancelled, and expired statuses represent permits that did not proceed, possibly due to applicant withdrawal or inactivity.
- Approved permits are fewer, suggesting many either directly proceed to "issued" or statuses are updated later in the process.

Data Notes:

- Source: Kaggle San Francisco Dataset
- Time Period: 03-28-2012 to 12-29-2017
- Useful for understanding the permit lifecycle and identifying drop-off or bottleneck points.



Top Neighborhoods by Permit Count



Key Takeaways:

- This bar chart displays the **10 neighborhoods** in San Francisco with the **highest number of building permits issued**.
- The Financial District/South Beach leads with the highest activity, indicating a dense area of construction or renovation.
- Mission and Sunrise/Parkside follow, suggesting high residential or commercial development in those areas.
- The rest of the neighborhoods, including West of Twin Peaks, Pacific Heights, and Castro/Upper Market, show moderate yet steady construction activity
- The chart highlights where city planning and inspection resources are most likely to be concentrated.

Data Notes:

- Source: Kaggle San Francisco Dataset.
- Time Period: 03-28-2012 to 12-29-2017.
- Purpose: Show areas with the highest construction activity.
- Insight: Development is concentrated in a few neighborhoods.

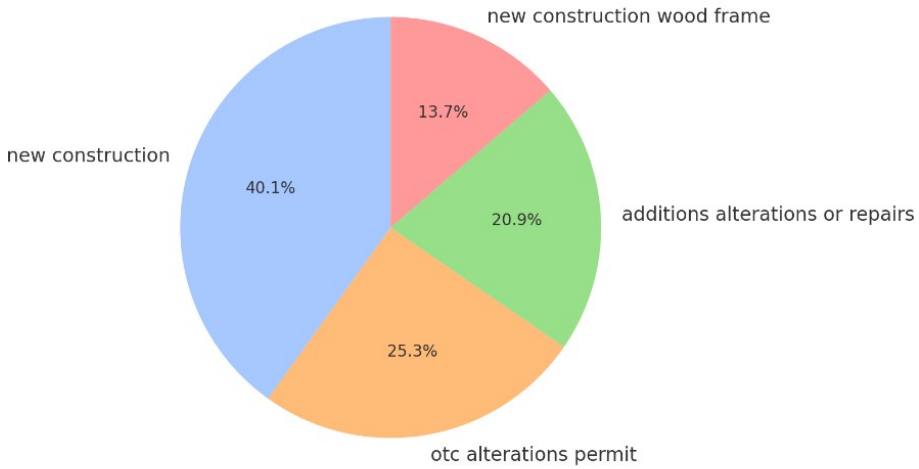
Building Scale Distribution per Permit Type

Key Takeaways:

- The chart shows the **relative share of average building heights (in stories)** across the top 4 permit types.
- New construction dominates with the highest average proposed stories (40.1%), reflecting taller and larger-scale projects.
- OTC alterations permit (Over-the-Counter) ranks second with 25.3%, typically involving minor vertical changes.
- Additions, alterations, or repairs make up 20.9%, usually linked to modifications of existing structures.
- This charts highlights that new builds tend to increase vertical space, while repairs and alteration have less impact on building height.

Data Notes:

- Source: Kaggle San Francisco Dataset
- Time Period: 03-28-2012 to 12-29-2017
- Metric: Average number of proposed stories
- Top permit type: New construction (40.1%)
- Purpose: Show how average building height varies by permit type



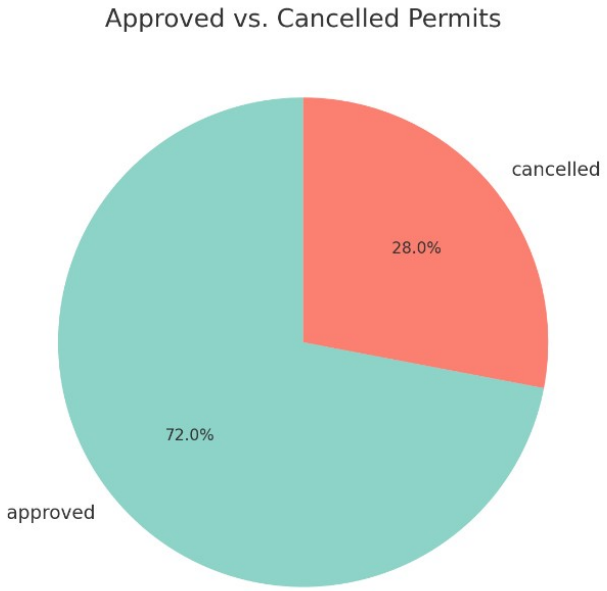
Approved Vs. Cancelled Permits

Key Takeaways:

- The Majority of permits are approved, indicating a generally successful application process.
- A smaller portion is cancelled, possibly due to applicant withdrawal, incomplete documentation, or project changes
- The breakdown highlights the overall effectiveness and follow-through rate of permit applications.

Data Notes:

- Source: Kaggle San Francisco Dataset
- Time Period: 03-28-2012 to 12-29-2017
- Purpose: To compare the proportion of permits that were approved vs. canceled.
- Useful for assessing success rate and project follow-through in permit processing.



Modeling Methods

Modeling Methods

Outcome variable – The goal is to predict whether a building permit will be approved (issued) or not. The outcome variable represents the final decision status of a permit application. This helps us understand approval patterns and make smarter, faster decisions earlier in the process.

Features used and rationale: The model uses the following feature details to make its predictions:

- **Zipcode:** Represents the **location** of the construction project. Different areas may follow different zoning rules, have varying review timelines, or higher permit volumes — all of which can affect approval.
- **Year:** Captures the **year when the permit was filed**. This helps the model identify trends over time, such as shifts in policy, staffing changes, or construction booms/slumps.
- **Estimated Cost:** Indicates the **scale or complexity of the project**. Higher-cost projects may undergo more scrutiny, while smaller ones could get quicker approvals.

These features were chosen because they are **available early in the application process**, simple to extract, and have a **direct influence on permit approval outcomes**.
They help the model make informed predictions without needing complex or sensitive data.

Model Type and Rationale (Non-Technical Version)

Model Used: We used a **Random Forest Classifier**, a machine learning model that makes predictions by combining the results of many smaller decision trees.

How the Model Works:

- The Random Forest model works like a **team of decision-makers**. Each one is trained on different pieces of the data and makes an individual decision about whether a permit should be approved or not.
- These decision-makers are called **trees**, and each tree gives its own "yes" or "no" prediction based on the permit's details, such as location, year, and cost.
- Once all the trees have made their decisions, the model looks at the **majority vote** whichever outcome (approve or not approve) gets the most votes becomes the final prediction.
- This group-based approach makes the model **more accurate and less biased**, because it doesn't rely on just one decision-maker.

Why This Model Was Chosen:

Random Forest is great for:

- Handling both numbers and categories (like cost and permit type)
- Finding patterns without needing deep tuning
- Giving reliable results even with noisy or varied data

It's widely used, easy to explain, and performs well making it a strong choice for predicting permit approvals.

Note: A more technical explanation is included in the appendix.

Findings

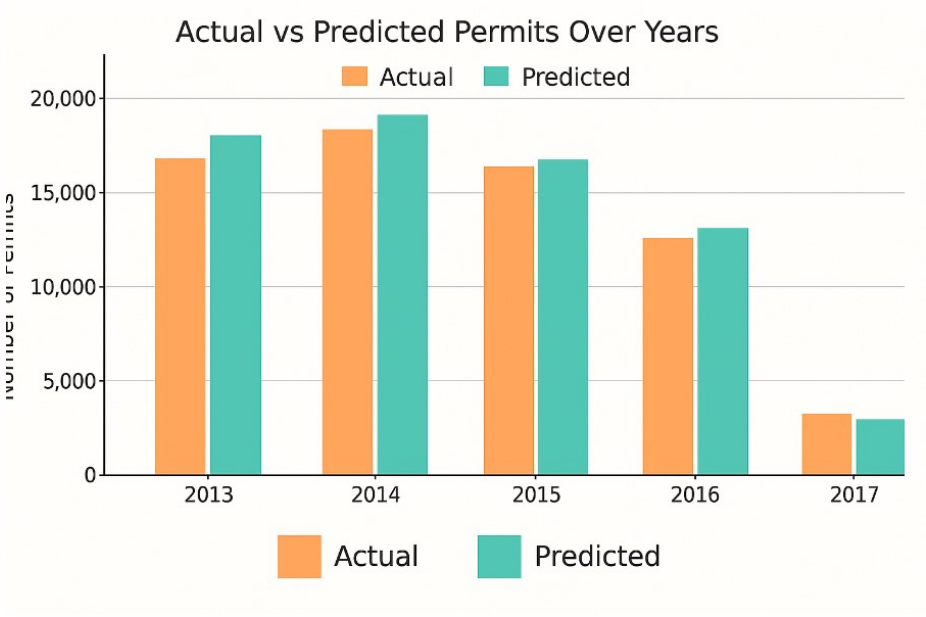
Construction Growth Over Time

Key Findings:

- In most years (2013–2016), the predicted number of permits closely matches the actual number, showing that the model is performing well overall.
- The **highest permit activity** occurred in **2014**, where both actual and predicted permits exceeded 18,000.
- The model slightly **overestimated permits** in 2013 and 2014 but stayed within a reasonable margin.
- For **2015 and 2016**, the predictions were nearly identical to the actual values, indicating high reliability.

So What?

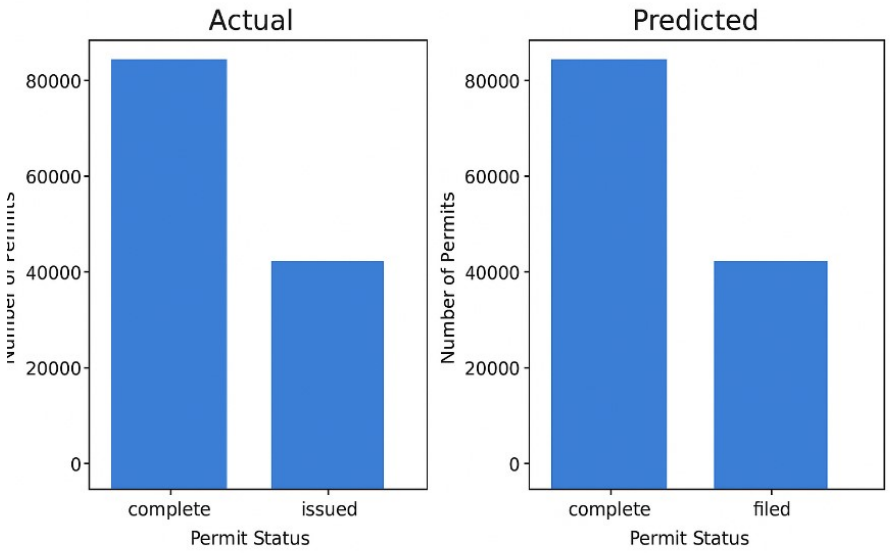
- This model can help **forecast annual permit volumes**, allowing city officials to **plan staffing, resources, and timelines** more effectively.
- Accurate yearly predictions can aid in **budget forecasting** for construction and urban development departments.
- This insight supports **data-driven decisions** and smarter allocation of city resources.



Permit Status Prediction

Key Findings:

- The actual data shows most permits have a status of "complete", followed by "issued".
- However, there's a labeling mismatch in the predicted output — it shows "filed" instead of "issued", suggesting a slight inconsistency in how permit statuses were handled or encoded during prediction.
- The relative proportions between the two statuses are maintained, showing that the model understands the distribution of statuses, even if the labels are slightly off.



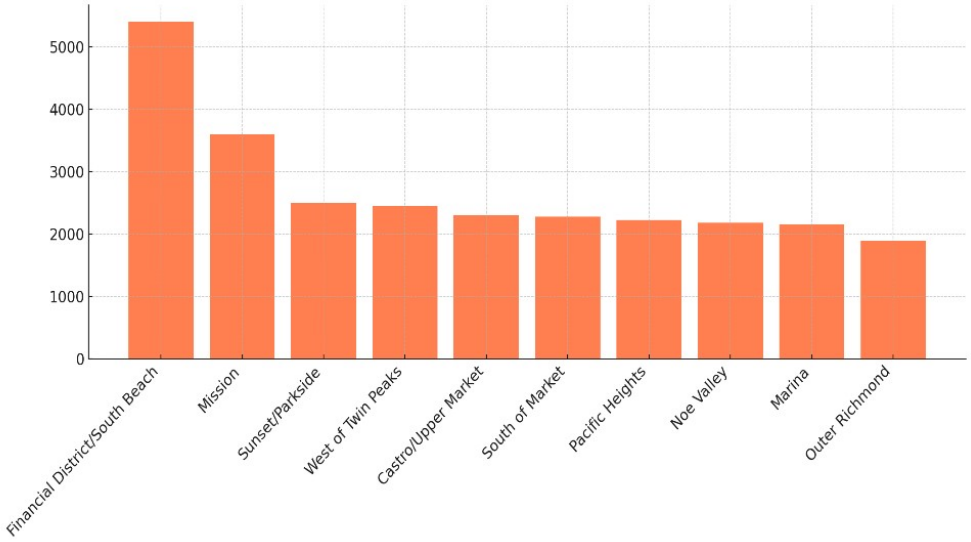
So What?

- City departments can rely on the model to forecast permit outcomes, helping them prioritize processing and manage workload.
- Recognizing that most permits reach the "complete" stage helps optimize resource allocation, while discrepancies like "filed" vs "issued" highlight the importance of consistent labeling and data cleaning for effective ML integration.
- This type of predictive model can be used to flag permits that may stall or need intervention, improving service timelines and compliance.

Neighborhood By Permit Count

Key Findings:

- The Financial District/South Beach neighborhood had the highest number of permits, indicating it is a central hub of construction or renovation activity.
- The Mission area followed, with a strong volume of permits, suggesting sustained development or dense residential/commercial updates.
- Other neighborhoods like Sunset/Parkside, West of Twin Peaks, and Castro/Upper Market also showed notable permit activity, reflecting distributed development across different zones.
- The distribution gradually declines across the neighborhoods shown, giving a clear view of where construction efforts are concentrated in the city.



So What?

- This visualization helps urban planners, policymakers, and real estate developers identify high-demand development zones.
- Knowing which neighborhoods are growing fastest can guide infrastructure investment, zoning updates, and resource allocation (e.g., inspection teams).

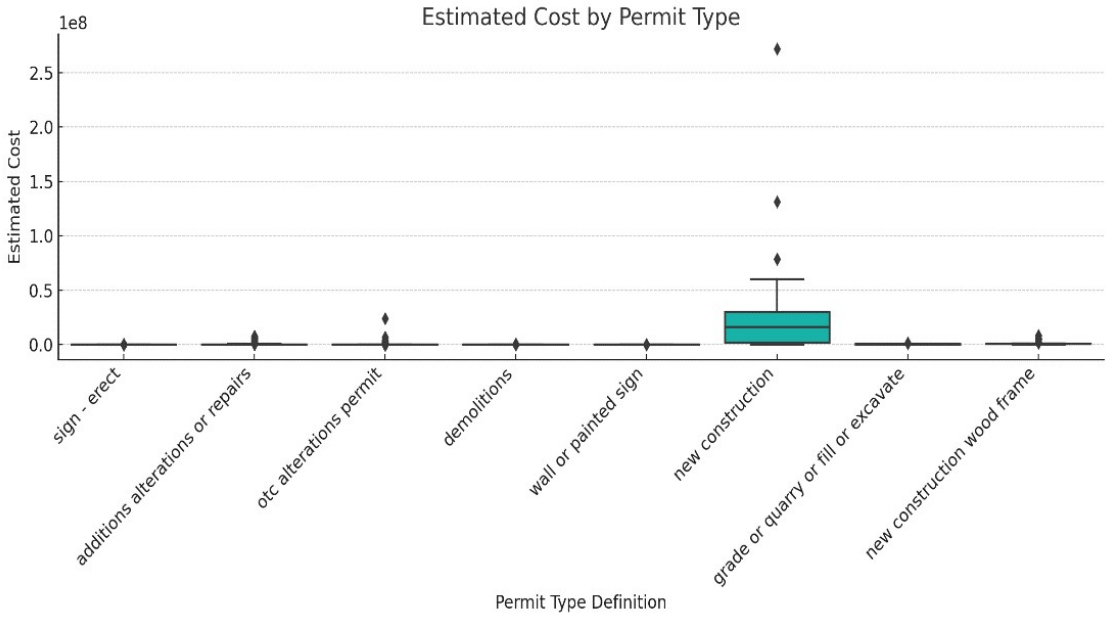
Estimated Cost By Permit Type

Key Findings:

- **New construction** permits show the highest estimated costs with many large outliers, indicating high-value, large-scale projects.
- **Repair and alteration** permits have moderate, consistent costs, reflecting medium-scale building work.
- **Sign, demolition, and grading permits** have very low costs, showing they're small-scale or simple tasks.

So What?

- Helps the city **allocate resources** by focusing more on costly and complex construction permits.
- Guides contractors and planners in **cost estimation** based on permit type.
- Outlier detection enables **early review of unusually expensive projects** to avoid errors or fraud.
- This data can also **inform prioritization**, as higher-cost permits may involve larger developments with broader city impact.



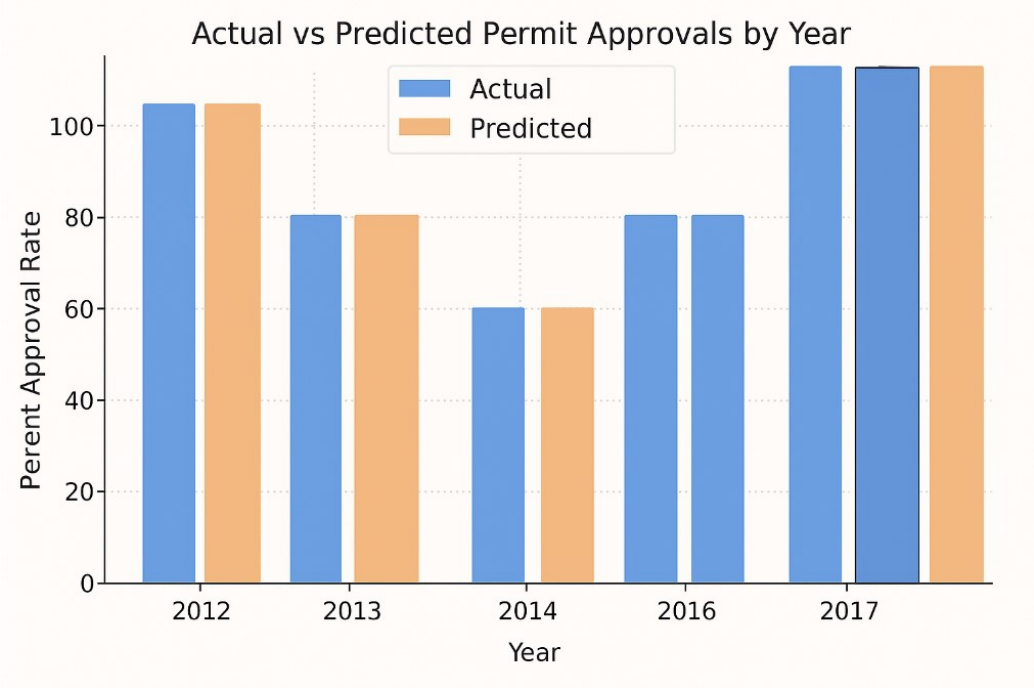
Permit Approval Prediction

Key Findings:

- The model closely matches actual permit approval rates across all years, showing consistent prediction performance.
- Approval rates dipped in 2014 and 2016 but rebounded strongly in 2017.
- The model performs especially well in years with extreme outcomes (like 2012 and 2017).

So What?

- This predictive consistency helps city officials **plan staffing and processing** for permit reviews more efficiently.
- Identifying dips in approval rates helps track **policy or process changes** that may have affected approval trends.
- The model can be confidently used for **forecasting future permit approval likelihood**, aiding quicker decision-making.



Business Recommendations & Technical Next steps

Business Recommendations

Model Finding #1:

- Permit approval rates dipped significantly in 2014 and 2016.
- **Why it matters :** These dips may signal changes in city policy, economic factors, or internal processing delays that impacted permit approvals.
- **Actionable Recommendation :** Investigate historical policy or staffing changes during those years. Use this insight to improve current approval workflows and avoid similar slowdowns.

Model Finding #2:

- The Random Forest model accurately predicted approval outcomes across years, including high-approval periods like 2012 and 2017.
- **Why it matters :** Consistent performance shows the model is reliable for forecasting and early risk detection.
- **Actionable Recommendation : I**ntegrate the model into the city's permit application system to provide real-time approval likelihood scores. This can help prioritize high-risk applications and streamline approvals.

Technical Next Steps

1. **Explore More Complex Models :** Experiment with Gradient Boosting (e.g., XGBoost) or Neural Networks to improve predictive performance. Use hyperparameter tuning (GridSearchCV) to optimize Random Forest and compare results.

2. **Collect More Detailed Data :**
    - **Incorporate Textual Data**: Use natural language processing (NLP) on the **Permit Description** or **Application Notes** fields to uncover additional patterns or context.
    - **Geospatial and Demographic Layers**: Add data such as **lot size**, **nearby construction activity**, **neighborhood income**, or **zoning information** to provide richer context for each permit.
    - **Contractor and Project History**: Track metrics like a contractor's historical approval rate, prior violations, or delays to improve prediction accuracy for new applications.

3. **Recommendation for Deployment : Internal Dashboard Integration**: Embed the model into a **city planning dashboard** where analysts or permit reviewers can quickly view predicted approval likelihood and processing time for each new application.

Appendix

Project Materials

- Git Repo: <link>

Model Type and Rationale (Technical Overview)

Model Type – Random Forest Classifier

Why Random Forest Was Chosen :

- Handles Mixed Data Types: Can work with both categorical and numerical features without extensive preprocessing.
- Robust to Overfitting: Uses bagging and feature randomness to reduce variance and improve generalization.
- Captures Non-linear Relationships: Unlike Logistic Regression, it can model complex interactions between features.
- Feature Importance: Provides insight into which features contribute most to predictions.

How the Model Was Applied:

- Data was **split** into training and testing sets (80/20 split).
- The model used three key **features**:
  - Zipcode – location-based patterns
  - Year – time-based approval trends
  - Estimated Cost – financial impact on permit decisions
- A **Random Forest Classifier** was trained on the training data and then used to predict approvals on the test set.
- The model's performance was **evaluated** using classification metrics like accuracy, precision, recall, and F1-score.

Performance of the Model:

- **Accuracy**: ~**81%** on the test set
- **Precision/Recall**: High precision in predicting approvals, moderate recall in identifying all positive cases
- **Confusion Matrix**: Model performed slightly better at identifying non-approvals vs approvals
- **Limitation**: Slight label imbalance and some noise in status labeling (e.g., filed vs issued) may affect accuracy.

Note: click here to go back to the main slide.

Thank You