Zillow Prize Challenge: Improving House Price Prediction

(SLIDE 1: Title Slide - Zillow Prize Challenge / Team ByteMe)

Good afternoon, everyone. Our team, **ByteMe**, is here today to present our solution for the Zillow Prize Challenge: **House Price Prediction Using Machine Learning**. Our core goal was clear: to significantly reduce the margin of error in Zillow's property valuations, thereby boosting user confidence in the Zestimate.

(SLIDE 2: Project Workflow / Methodology Overview)

Our approach followed a rigorous, four-step data science workflow. We began with extensive data cleaning and feature engineering, creating temporal and geospatial variables crucial for real estate. This foundation allowed us to move into model selection, where we prioritized algorithms that offered both high performance and interpretability. The culmination of this work is the model we're presenting today, which provides a robust and explainable valuation platform.

(SLIDE 3: Model Architecture and Selection)

For this critical task, we selected the **Random Forest Regressor**.

Why Random Forest? It offers an optimal balance of key attributes:

1. **Robustness:** It handles outliers and missing data exceptionally well, which is essential in messy real estate datasets.
2. **Performance:** It captures complex, non-linear relationships without manual fine-tuning.
3. **Interpretability:** Crucially, it provides clear, quantifiable feature importance rankings, allowing us to explain *why* a prediction was made.

(SLIDE 4: Model Performance Metrics)

Let's look at the results. We evaluated our model using the **Mean Absolute Error (MAE)** to measure the average magnitude of our log-error predictions.

Our results demonstrate strong performance:

- **Training MAE:** 0.068252
- **Testing MAE:** 0.070266

The critical insight here is the minimal difference—only 0.002—between our training and testing MAE. This confirms that our model has achieved excellent generalization. We are not overfitting the training data; the model performs just as well on unseen, real-world data, making it highly reliable for deployment.

(SLIDE 5: Feature Importance Analysis)

This is arguably the most valuable slide for the business. Understanding the drivers behind value is how we build trust.

The chart shows the relative importance of each feature, and the results provide powerful validation of real estate domain knowledge:

- The strongest single predictor, at nearly 14%, is the **Tax Assessment Value**. This shows that Zillow's existing internal or external valuations are highly informative when correcting for log-error.
- Next are the fundamental drivers: **Living Area Square Footage** and **Lot Size** confirm that physical size remains a core factor.
- Finally, **Geographic Factors**—specifically Latitude, Longitude, and Zip Code—are extremely important, capturing neighborhood-level effects, school districts, and local market demand trends.

The model is prioritizing exactly what real estate professionals would expect, making it highly trustworthy.

(SLIDE 6: Key Insights and Business Value)

So, what is the business impact of this model?

The primary business value is the direct **improvement of Zestimate accuracy**. By accurately predicting and correcting log-errors, we can deliver more precise valuations. This translates directly into **increased user trust** in Zillow's platform, reducing hesitation from buyers, sellers, and agents who rely on accurate estimates.

(SLIDE 7: Conclusion & Future Enhancements)

In conclusion, our approach using the Random Forest Regressor provides a powerful, robust, and *interpretable* foundation for refining Zillow's valuation services, achieving a strong MAE of 0.070266 on our test set.

Looking forward, we see three key areas for enhancement:

1. **External Data Enrichment:** Integrating external data like local economic indicators, school ratings, and recent neighborhood crime data will add vital context.
2. **Algorithm Experimentation:** While Random Forest is excellent, we plan to explore advanced, high-performance gradient boosting algorithms like XGBoost and LightGBM for potential incremental gains.
3. **Hyperparameter Tuning:** A dedicated grid search with cross-validation will ensure we are extracting the absolute maximum performance from our current architecture.

Thank you for your time. We are happy to take any questions.

Team ByteMe, November 2025.