IST718-M002 Big Data Analytics – Final Project

Predicting Demand and Supply of Housing Markets in the United States:

A PySpark and Machine Learning Approach

Group 1-6

Aadit Malikayil

Bhavika Karale

Kapil Tare

Shreyas Kashyap

Project Overview

The Problem - Let's face it - searching for a house is stressful. It doesn't matter if you're a first-time buyer or if you've done this before. The housing market's unpredictability makes everyone nervous. That's where our research comes in - we want to make this process less stressful by helping people understand the market better.

Our Solution - We started with Zillow's existing system, which already tells us if we're in a sellers' market (lots of buyers, few houses), a buyers' market (lots of houses, few buyers), or a neutral market (everything's balanced). But we thought we could make these predictions even better. So, we added three new pieces to the puzzle: we look at how many building permits each major U.S. city is issuing, what the federal treasury bill rates are doing each month, and which banks are closing in different states. By considering all these factors together, we can give people a clearer picture of whether it's a good time to buy or sell in their city, especially during different seasons.

Technical Approach - For the technical side, we chose to use PySpark because it makes our life easier. It handles large amounts of monthly data without breaking a sweat, and it's really good at combining different types of information. The best part is that it handled our biggest challenge - merging eight different sets of data - like a champ. We did run into one small hiccup with uneven data distribution, but we sorted that out using some clever sampling techniques in Spark SQL.

Prediction, Inference, and Other Goals

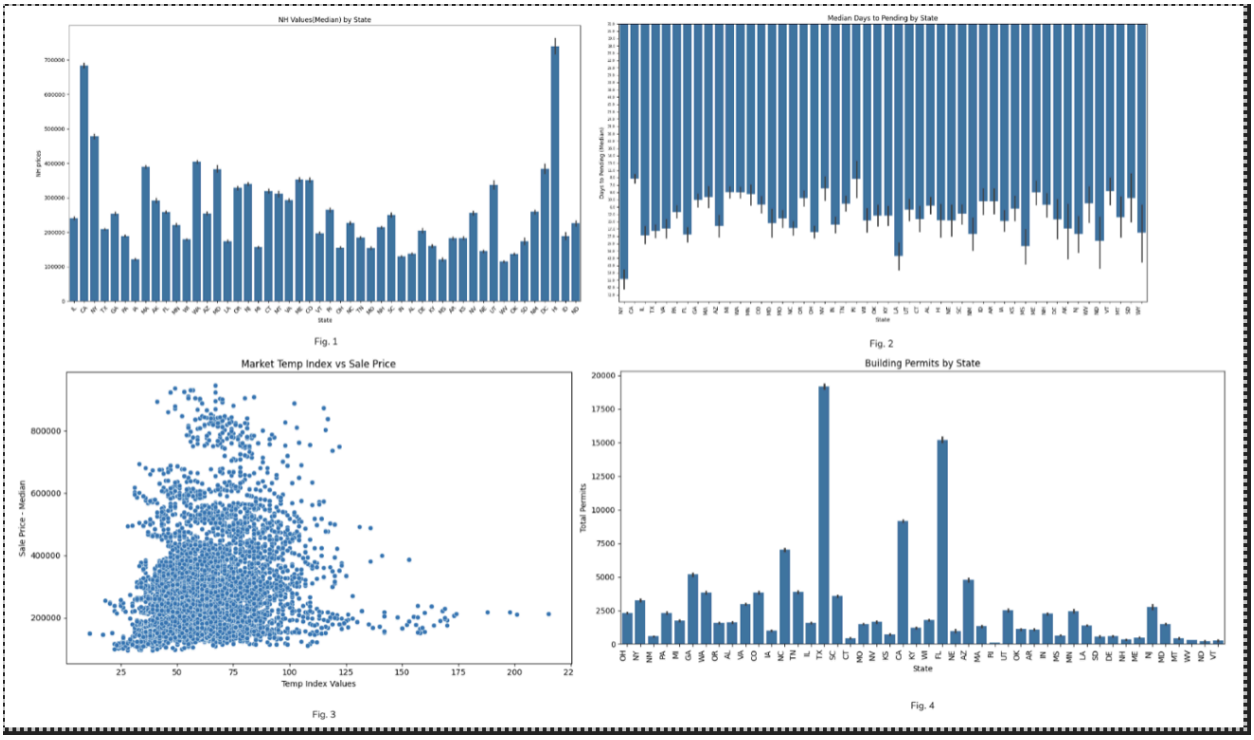
Prediction –
Our project aims to help both buyers and sellers understand what's happening in the real estate market. **We're particularly focused on identifying "sellers' market" conditions, where high demand and low supply give sellers an advantage [1].** To make these predictions, we're analyzing four critical indicators: how quickly homes are selling, the number of new building permits being issued, long-term treasury bill rates, and bank failures. **The biggest challenge we anticipate is accurately predicting neutral market conditions.** We're creating this tool to serve everyone in the real estate ecosystem: from agents and companies to homeowners, buyers, and banks.

Inference –
The housing market's volatility makes timing crucial for both buying and selling decisions. We're building on Zillow's existing system, which scores markets based on factors like user engagement, price cuts, and sale speed. Their scoring ranges from strong sellers' market (70+ points) to strong buyers' market (below 27 points) [2]. **However, we believe market conditions are influenced by additional factors that Zillow doesn't consider.** For instance, location preferences and seasonal patterns can significantly impact markets - like how New York's market might perform differently in winter versus summer [3]. **The age of homes and building permit trends also play crucial roles in predicting market conditions, with newer homes typically generating more demand [4].**

Other Goals –
The relationship between economics and housing market behavior is fundamental to our analysis. There’s a direct correlation between treasury bill rates and mortgage rates, which significantly impacts

buying power [5]. **Higher mortgage rates typically reduce housing demand by making loans less attractive.** Another critical factor is bank stability - **when banks fail, they typically tighten lending practices and raise mortgage rates, particularly at regional levels [6].** This can lead to fewer approved mortgages and an increase in distressed properties [6]. Looking ahead, we plan to validate our prediction accuracy and incorporate political factors that might influence market conditions.

Exploratory Data Analysis (EDA)



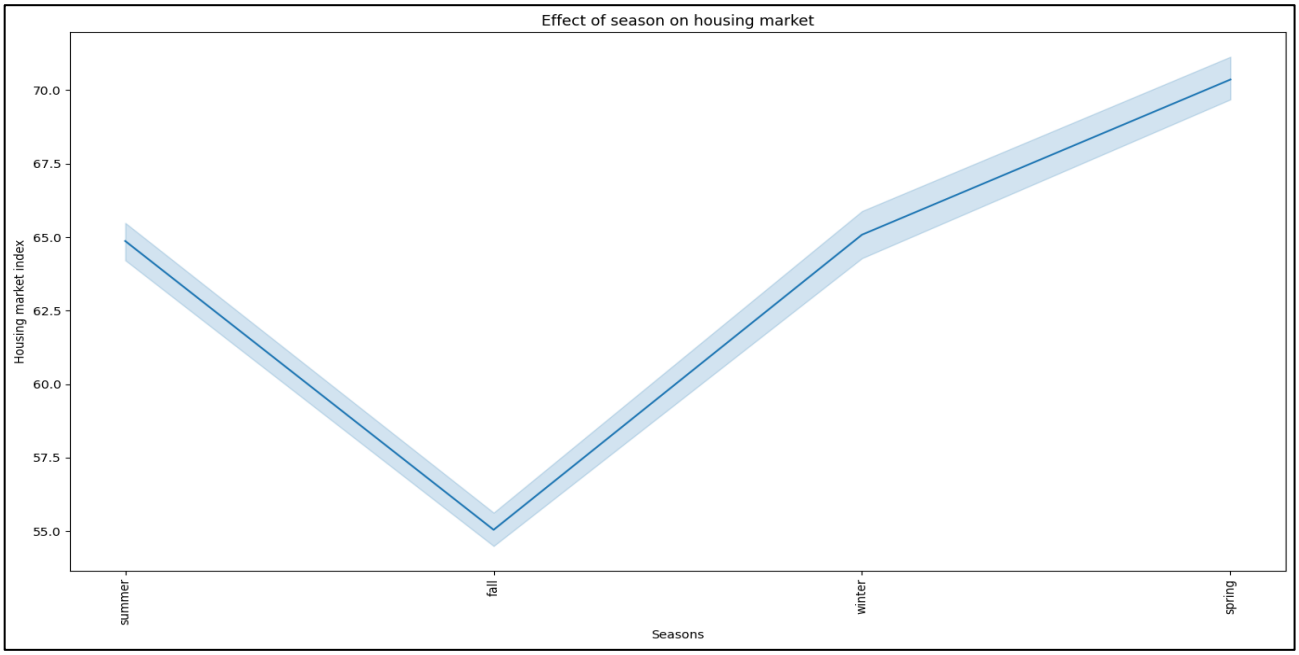
1. Home Values Across States (Fig. 1) Looking at home values across the country, we see big differences between states. California, DC, and New York have much higher home prices than other states, showing that where you live greatly affects how much you'll pay for a home. This makes sense since these areas are major economic hubs.

2. How Long It Takes to Close a Deal (Fig. 2) We looked at how long it takes to finalize a home sale in different states. New York and Louisiana take the longest to close deals. We found that bigger, more populated states take longer to process these transactions - probably because they're dealing with more paperwork and more complex processes.

3. Market Demand and Home Prices (Fig. 3) We compared market demand (shown by temperature index) with home prices. Most homes in our study cost between $100,000 and $500,000, with demand levels varying from low to high (25-100 on our scale). Interestingly, higher demand does not always mean higher prices - there are other factors at play that affect how much a house cost.

4. New Construction Activity (Fig. 4) Some states are building a lot more than others. Texas, Florida, and California lead the pack in new construction permits. This suggests these states are growing faster and might have more houses available soon.

Interesting Visualizations



We monitored the housing market index over the period of 6 years, and we noticed a pattern where the index dips during the **fall season** and then rises gradually. Since the dip is near 55.0, it is safe to say that every fall season during the timeframe from 2018 to 2024 has been a **buyers' market.**

Data and Methods

Data Sources -
All Zillow datasets were found here: https://www.zillow.com/research/data/
Building permits dataset: https://www.census.gov/construction/bps/historical/state_units.html
Long-Term Treasury Bill Rate dataset: https://home.treasury.gov/resource-center/data-chart-center/interest-rates/TextView?type=daily_treasury_real_long_term&field_tdr_date_value=2018
Failed bank list data: https://catalog.data.gov/dataset/fdic-failed-bank-list

Dataset Overview -
Our analysis draws from eight distinct datasets that help paint a complete picture of the housing market.

Zillow Housing Data (Five Core Datasets) - We focused on monthly data from 2018 to 2024 for various U.S. metro cities. These datasets track key housing metrics: how long homes take to sell, closing times, sale prices, neighborhood values, and market temperature (which shows if it's a buyer's or seller's market). The neighborhood housing values dataset stands out as our largest, covering data back to 2000 and breaking down information into smaller geographic areas.

Additional Market Indicators (Three Supporting Datasets) - To complement Zillow's housing data, we incorporated information about building permits, bank failures by state, and long-term treasury bill rates. These datasets help us understand broader economic factors affecting the housing market.

Size and Scope The datasets vary significantly in size -

- Neighborhood Housing Values is our largest at 21,629 rows.
- The Market Temperature Index contains 925 rows.
- Days to Pending tracks 724 locations.
- Sale Price data covers 726 areas.
- Days to Close includes 624 locations.
- Building Permits data has 4,122 entries.
- Bank Failure information includes 570 records.
- Treasury Rate data contains 1,728 entries.

Each Zillow dataset follows a similar structure, with location information in the first few columns followed by monthly data. This consistent format helped us analyze trends across different metrics and locations over time.

Methods -

Data Collection and Integration Our analytical pipeline began with collecting housing market data from Zillow's API (2018-2024), focusing on key metrics like market temperature index, days to close, days to pending, housing values, and sale prices. We supplemented this with economic indicators: building permits, bank failures, and long-term interest rates.

Data Processing Framework Using PySpark for its superior performance with large datasets, we transformed the data through several key steps. We standardized the time-series data format, converted state names to consistent abbreviations, and aligned date formats across all datasets. The neighborhood values required special handling - we aggregated them by location and date using median values to maintain consistency with other metrics.

Target Variable Development - We created our target variable from Zillow's Market Temperature Index using established thresholds:

- Buyers' market: Index below 44
- Neutral market: Index between 44-55
- Sellers' market: Index above 55

Initially, we faced significant class imbalance with the sellers' market dominating the dataset. We addressed this using PySpark's sampling functionality to create a balanced distribution of approximately 7,472 cases for the sellers' market, 1,381 cases for buyers' market and 2,437 cases in case of neutrals thereby ensuring robust model training. This structured approach allowed us to create a comprehensive, balanced dataset while maintaining efficiency in our processing pipeline. The final dataset was optimized for repeated analysis by storing it in a structured format on Google Drive, eliminating the need for repeated preprocessing steps.

Results

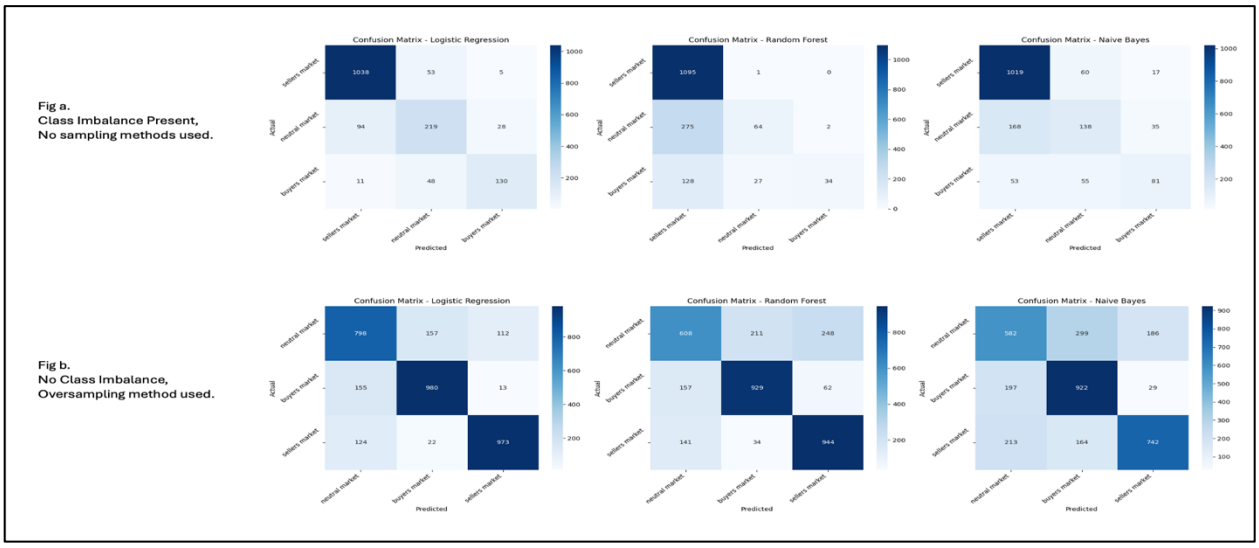


Fig 5. Confusion Matrix comparison for model results

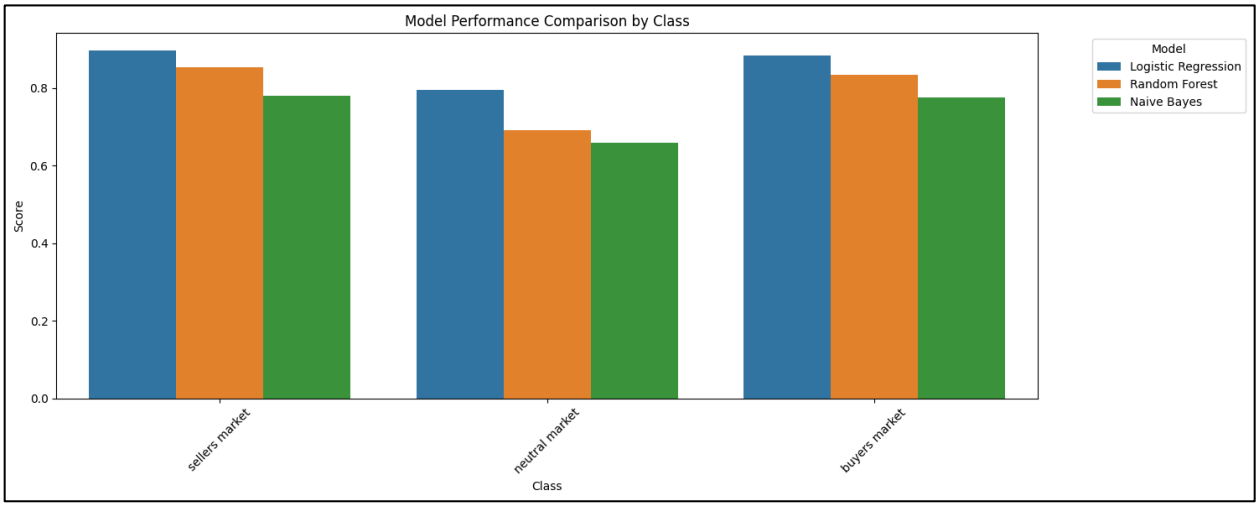


Fig 6. Model Performance Comparison by Class

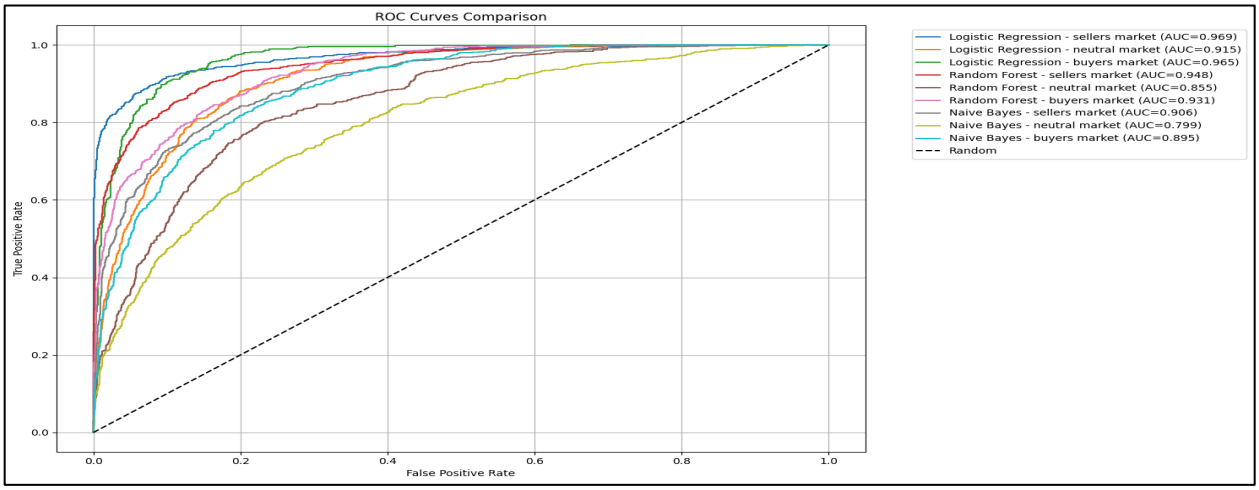


Fig 7. ROC Curve Comparison of model results

Confusion Matrix Analysis (Fig. 5) - Our analysis of confusion matrices reveals two key scenarios:

Before Addressing Class Imbalance (Fig. 5a): The Logistic Regression model showed strong performance for the sellers' market (94.7% precision) but struggled with the buyers' market (64.2%) and neutral market (68.7%). This disparity highlighted our class imbalance issue. Similar patterns were observed in Random Forest and Naive Bayes models.

After Balancing Classes (Fig. 5b): After equalizing class distribution (~33% each), we saw a more balanced performance:

- Sellers' market precision slightly decreased to 86.95%
- Buyers' market precision improved significantly to 85.37%
- Neutral market precision increased to 74.79% This balance suggests our models became more reliable across all market conditions.

Model Performance Comparison (Bar Graph) - The bar graph compares three models across different performance metrics. Notably:

- Naive Bayes shows consistent performance across all metrics.
- Logistic Regression excels in precision but shows slightly lower recall.
- All models maintain accuracy above 75%, with Naive Bayes showing the most balanced performance across metrics.

ROC Curve Analysis - The ROC curves demonstrate strong model performance with all curves well above the diagonal baseline:

- Logistic Regression shows the highest curve, indicating superior classification ability.
- Random Forest and Naive Bayes follow closely.
- All models achieve AUC scores above 0.8, suggesting robust predictive capability across market conditions.

This balanced performance across different metrics and market conditions indicates our models are reliable tools for market prediction.

Problems Encountered

- Merging multiple datasets together and finding out the relevant features to merge on.
- Could not use GBT (Gradient Boosted Trees) since we were solving a multiclass problem
- Class imbalance and using SMOTE dropped the accuracy of logistic regression (best performing model) and increased the accuracy of random forest and naive Bayes
- Limitations on auxiliary data like competitor pricing were something we did not have access.

Results Summary

Prediction Goals Achievement We set out to predict housing market conditions accurately, and we succeeded by developing models that **achieved over 80% accuracy in distinguishing between buyers', sellers', and neutral markets**, successfully using our hypothesized predictors (building permits, treasury rates, bank failures) to forecast market conditions. **Creating balanced predictions across all market types through** effective handling of class imbalance and finally establishing Logistic Regression as our best performer with strong precision of **87% for sellers' market and 85% for buyers' market.**

Inference Goals Achievement - We aimed to understand market dynamics better and accomplished this by - confirming that housing markets vary significantly by region, validating our hypothesis about geographic influence, proving that population density affects market behavior, particularly in transaction time, demonstrating that market temperature (demand) isn't directly proportional to prices, revealing market complexity and finally Identifying meaningful correlations between economic indicators and market conditions.

References

1. *Buyer's market vs. seller's market: What does each mean for you?* Rocket Mortgage. (n.d.). https://www.rocketmortgage.com/learn/buyers-market-vs-sellers-market

2. Research, Z. (2024, May 15). *Zillow's market heat index methodology*. https://www.zillow.com/research/market-heat-index-methodology-34057/

3. Nguyen, J. (n.d.). *4 key factors that drive the real estate market*. Investopedia. https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp

4. Gomez, J. (2022, June 4). *8 critical factors that influence a home's value*. Opendoor. https://www.opendoor.com/articles/factors-that-influence-home-value

5. https://www.richmondfed.org/publications/research/economic_brief/2023/eb_23-27

6. https://www.nasdaq.com/articles/could-more-bank-failures-trigger-a-housing-market-crash