Real Estate Property Recommendation Engine

Motivation

- Recommend new locations to buy and invest properties based on broad regional real estate factors.
- Better opportunities for both individual buyers and investors throughs insights from the engine.
- Helps real estate investors with data driven decisions from the interactive search tool.

Current Approach & Innovation

- It will be a major improvement compared to current micro home-based search.
- Predictive regional models based on:
    - Broad regional real estate factors: inventory, median-days-to-close, Days on the market, etc.
    - Macro-economic factors such as interest rates, inflation, consumer confidence, etc.
- Better insights for buyers, sellers, lenders and government agencies.

Data & Exploratory Analysis

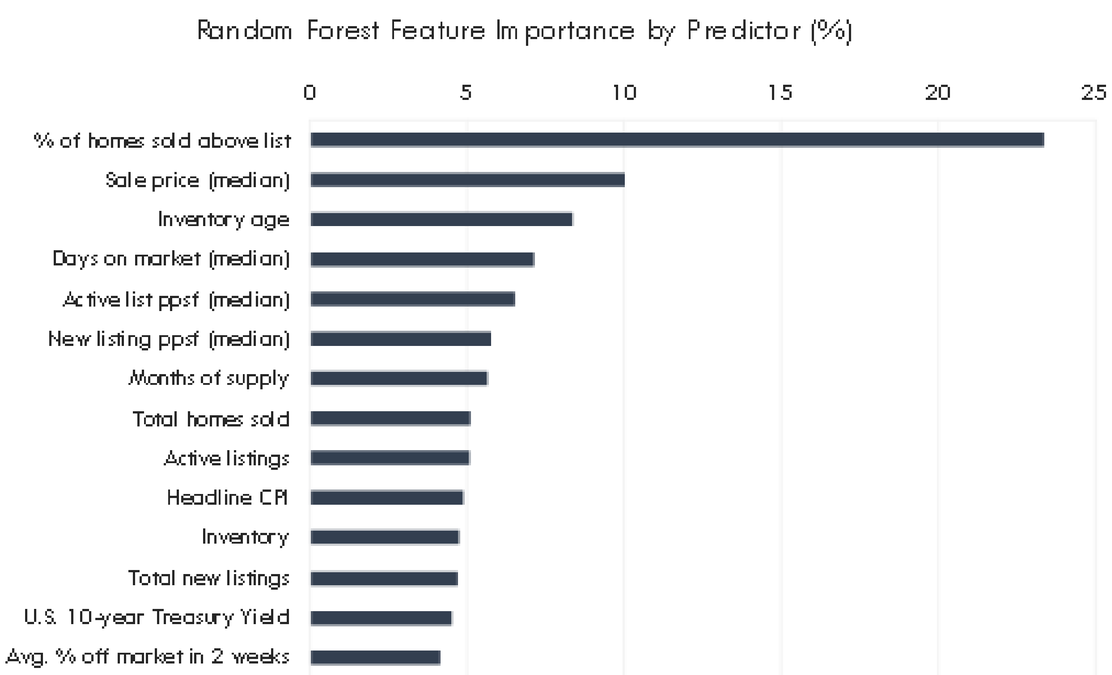
Data:

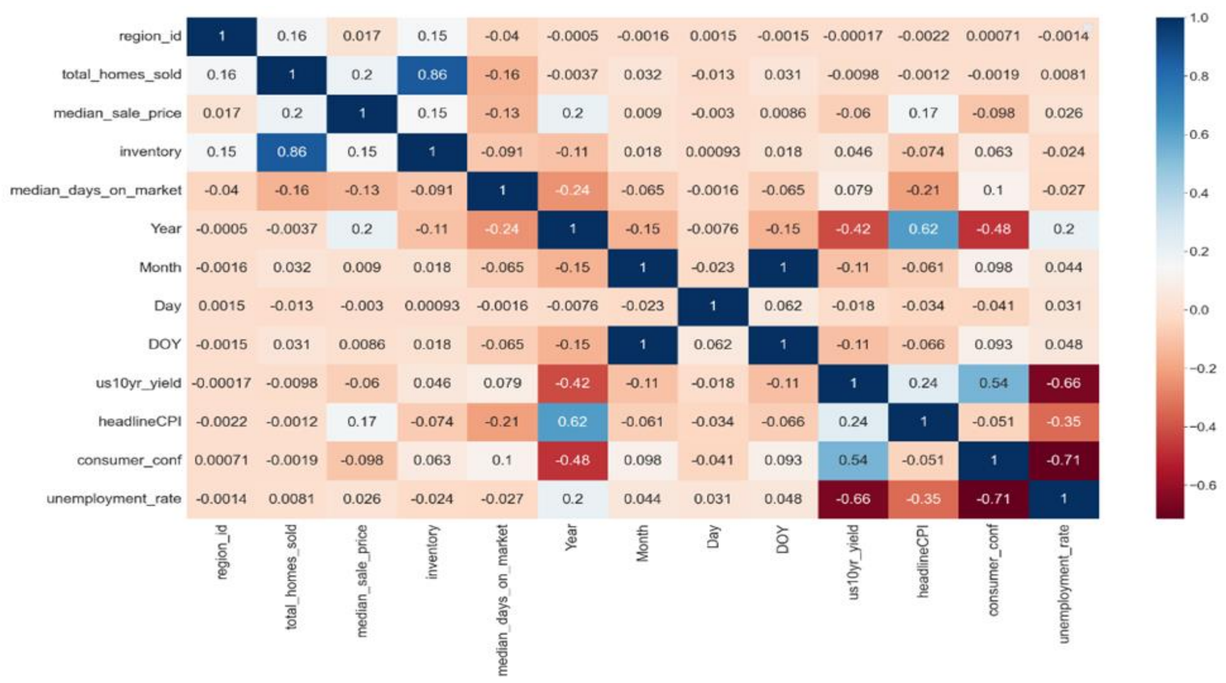
- Redfin's "Weekly Housing Market Data" to evaluate and train our models.
    - Source: Redfin Data Center (https://www.redfin.com/news/data-center/)
    - Raw Data size: 2.93 GB, # of rows: 3, 472, 095, feature columns: 96.
- CPI – Macro economic factor used in the model. Source: FRED API

Data Preparation and assumptions:

- Restricted our study to regions at least >= 150 data records.
- Excluded columns with >= 25% of null or nan values.
- Points with median days on market greater or equal to 1095 were removed – outliers
- In the Redfin data is available for different durations. For our models, we leveraged duration 1 weeks.

Feature Importance for Random Forest and Log Regression Correlation Plots:





Modeling Algorithms, Training, & Test Data

- There are mainly **two features** in our project:
    - Provide Heatmap of expensive vs cheap property locations using the classification. For this, we leveraged avg-sale-to-list ratio. For this feature, we experimented with Decision trees, Random Forests, and XGBoost.
    - Second feature is to predict the probability of being able to sell a property within a certain period. For this feature, we employed Logistic Regression model.
- **Random Forest** is a meta-classifier that trains a number of decision tree classifiers on various sub-samples of the dataset and uses the average of the prediction to improve the prediction accuracy and avoid overfitting.
- **Logistic Regression** is a statistical model that models the probability of an event taking place by having the log-odds for the event.
    - Every city and county has its own model
    - Dependent variable: Probability of exceeding days on the market cutoff
    - Independent variable: median_house_price, days_on_market_cutoff, headline CPI, season
    - Train and Test data assumptions:
        - Data is split based on year as this is a time series split.
        - Training data: 2017 to 2022 and Test data: 2022 to 2023
        - Headline CPI for testing: 6.51

Results Discussion & Evaluation Metrics

After building the models described in the previous section, we evaluated the models with test data and computed various metrics as shown below:

For probability model, metrics are computed by taking the average of 2905 regional logistic models across the country. We have built regional models for each region i.e. city / metro region in the input raw dataset.

	Random Forest Classifier (%)	Logistic Regression Model (Avg %)	Explanation
Accuracy	83.0	55.7	Random forest model has a high accuracy and logistic regression model does a fair job.
Precision	83.0	63.9	Precision is high for identifying if an area is cheap or expensive but for predicting if property stays on market above a cut off date is moderate.
Recall	81.0	84.1	Both the Random forest and logistic regression models have high recall values
F1-score	82.0	61.5	Score is moderately good. Higher the better. This is particularly useful for logistic regression as we may end up with an imbalanced dataset due to different cut off dates

- Using the Redfin's regional level Weekly Housing market data and macro-economic factor CPI index, we built a Random Forest model to create the affordability index.
- Built a Logistic regression model for each region to predict the probability of exceeding a given number of days a property sits on the market.
- Analyzed the results with various metrics as explained in the results section.
- Built the Visual presentation of our tool using Tableau.

Limitations and Future Work

- Due to a lack of access to the micro-level home level data, we could not build the complete recommendation engine.
- In our dataset, we used the Redfin data from 2017 to early 2023. We believe that due to the unusual pandemic era boom in home prices and inflation, there may be some unusual trends in our test data.
- Due to time constraints, we built the visualization as Tableau dashboard. Web application will be part of our future work.