Housing Market Decision Tree Classification

DTSA 5509: Introduction to Machine Learning-Supervised Learning

Author: Andrew Whittenbarger

Agenda

- Problem
- Data Description and Cleaning
- Exploratory Data Analysis
- Modeling
- Results

Problem: When is the right time to buy or sell a house?

Buyers Market: Market conditions favor the buyer. The buyer is more likely to get a home below the listing price.

Sellers Market: Market conditions favor the seller. The seller is more likely to get at or above the listing price for their home.

Goal: Build a classification model to help prospective buyers or sellers determine if now is the right time to enter the market.

Data Description and Cleaning

- The data analyzed is a combination of 3 datasets from Zillow

Dataset	Months (Columns)	Metros (Rows)	Missing Values
Days to Pending (DTP)	63	775	16829
Inventory (INV)	63	898	1254
Sale-to-List Price Ratio (STL)	62	494	6550

- All datasets also contained 5 categorical columns: RegionID, SizeRank, RegionName, RegionType, and StateName

Data Description and Cleaning (Cont.)

- March 2023 Data was dropped since it wasn't contained in all datasets
- Metro Areas not contained in all 3 datasets were also removed
- "RegionID" and "RegionType" were removed because it was known they would not be used
- STL Ratio was converted to a categorical variable: "Market_Type"
  - STL<1 was considered a buyer's market (-1)
  - STL>=1 was considered a seller's market (1)

Data Description and Cleaning (Cont.)

- An attempt was made to impute missing values by averaging neighboring values
- When the datasets were consolidated rows that still had missing values were dropped
- Final data frame contained 25,884 samples

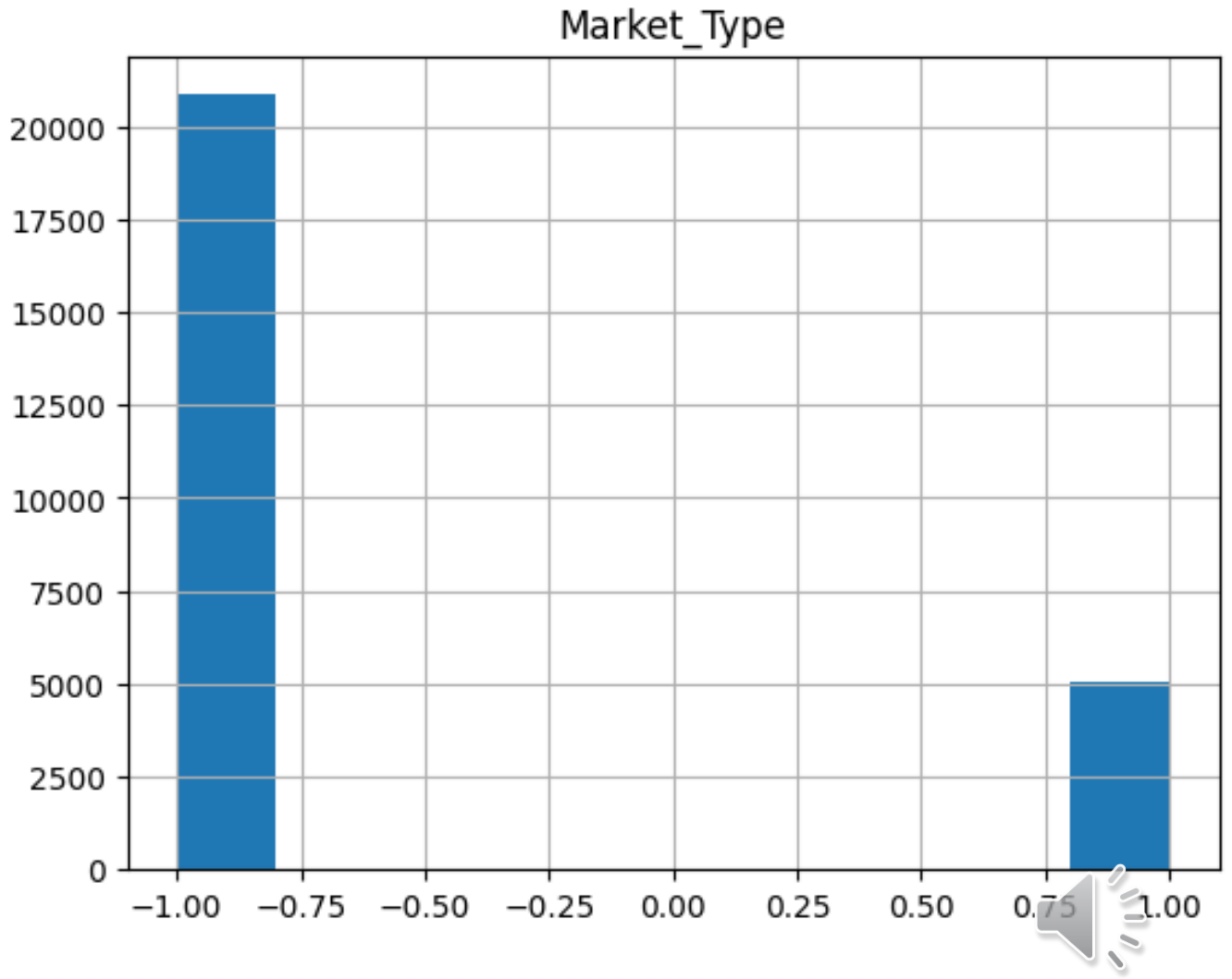
SizeRank	Metro	State	DTP	INV	Market_Type
1	New York, NY	NY	73	66530	-1

(Example of row from final data frame)

Exploratory Data Analysis

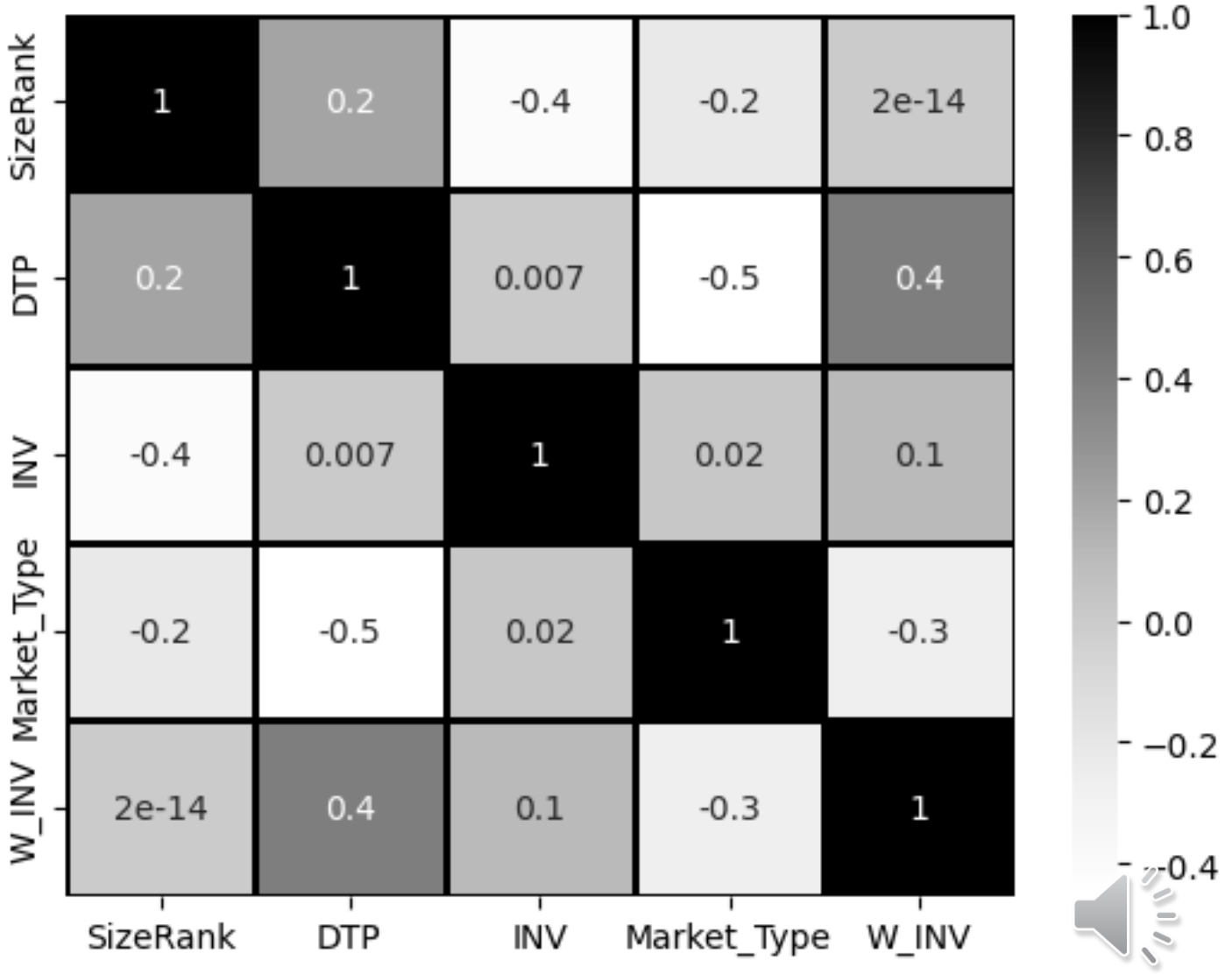
Key Takeaways

1. Most samples are from buyer's markets, so the dataset is imbalanced
2. Inventory should be weighted to reflect market averages (raw number not meaningful)



Exploratory Data Analysis (Cont.)

correlation coefficients after creating a weighted inventory

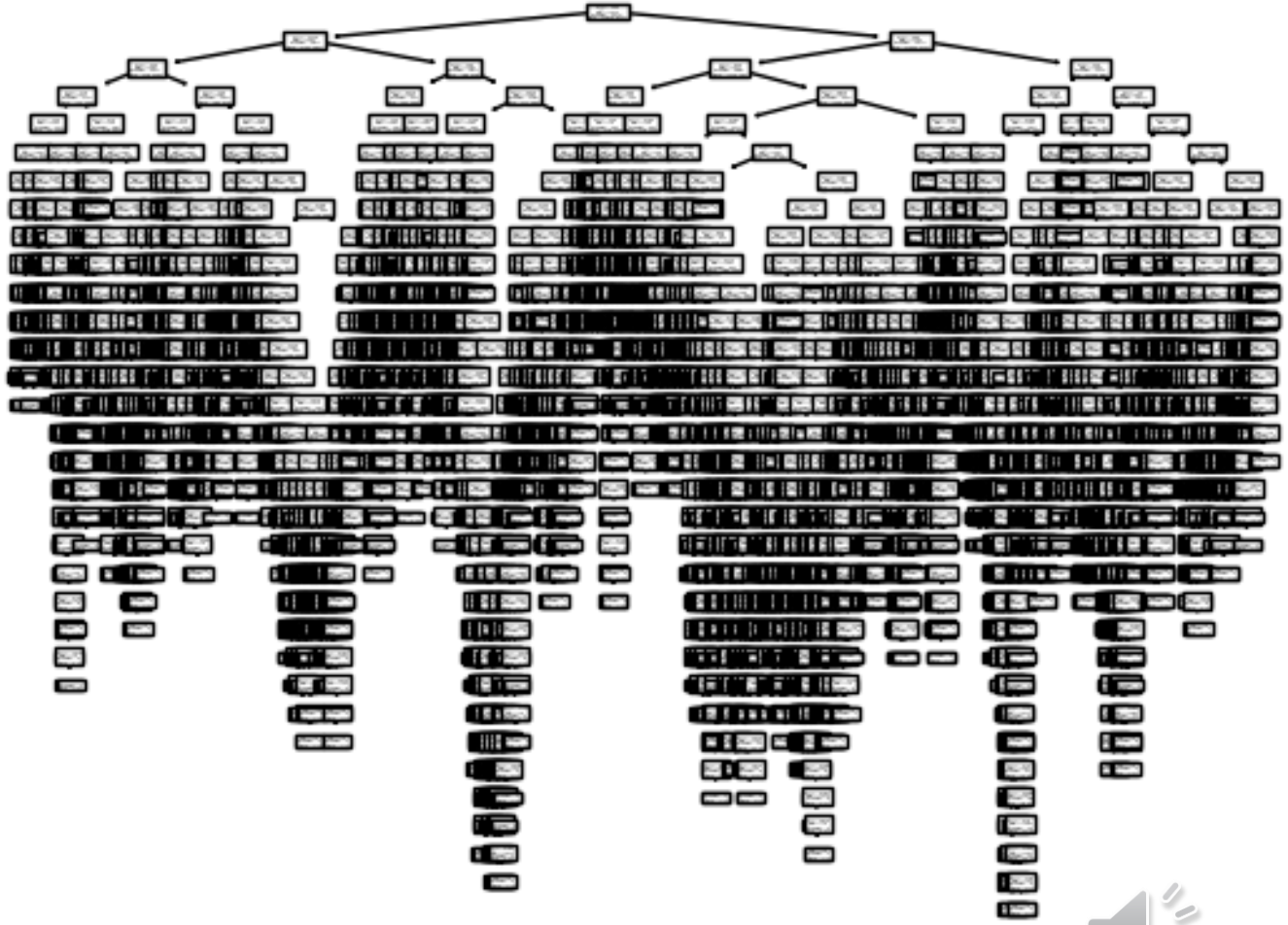


Data Modeling

- Features Selected: Days to pending, SizeRank, and the weighted inventory (W_INV)
- Response Variable: Market_Type
- Data was split into a training and test set using an 80/20 split ratio
  - Data stratified to ensure the training and test dataset contained the same ratio of Market_Types

Data Modeling: Simple Decision Tree

- Decision tree Built using the Sklean repository
- Depth: 32
- Nodes: 5199
- Feature Use: Days to Pending (48%), SizeRank (28%), Weighted Inventory (24%)
- Likely some overfitting
  - Will use ensembling and hyperparameter tuning to address



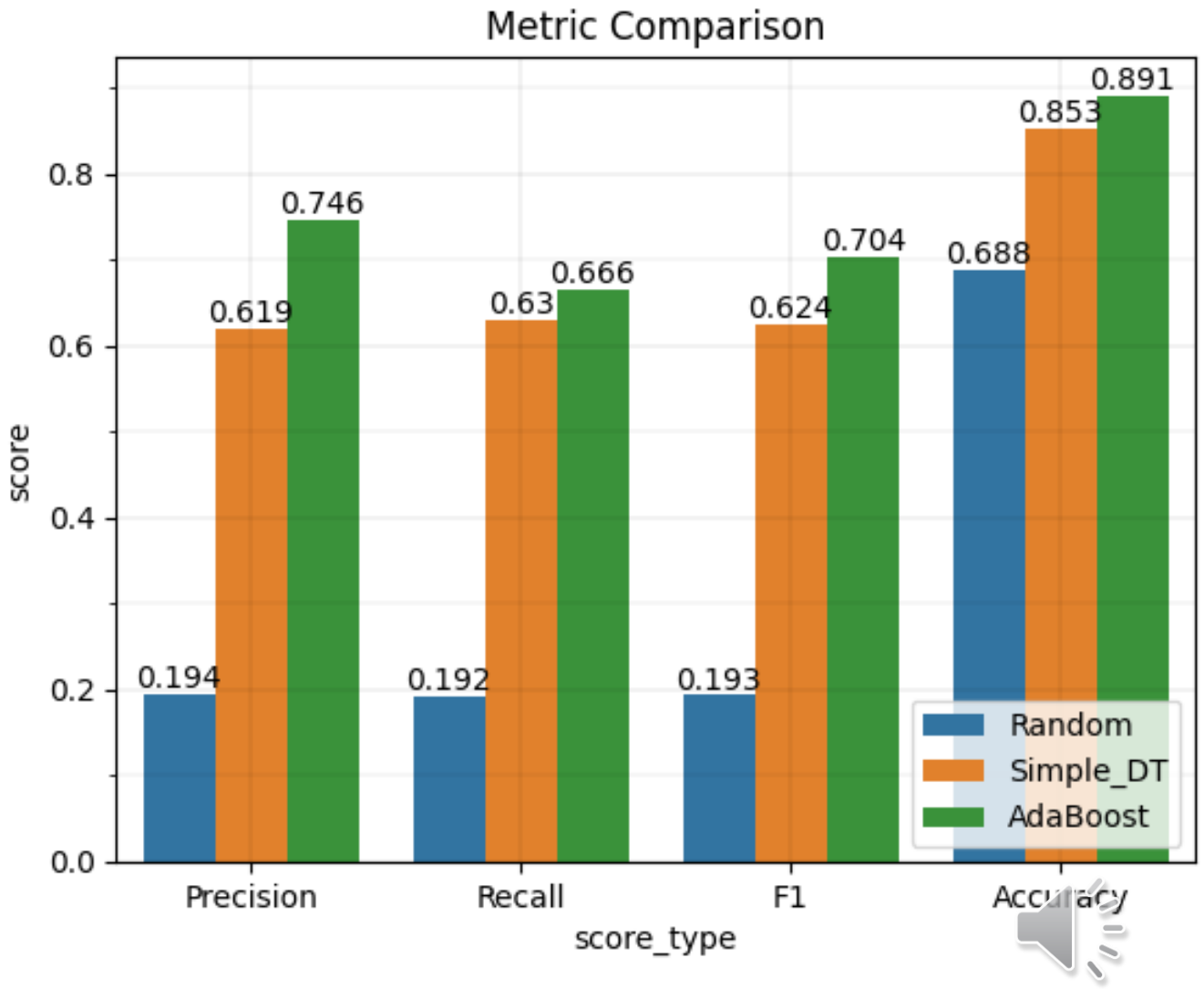
Data Modeling: AdaBoost Decision Tree

- GridsearchCV used for cross validation and hyperparameter tuning
- F1 Score Prioritized as scoring method over accuracy since the dataset is imbalanced
- Feature Use: SizeRank (43%), Weighted Inventory (36%), Days to Pending (21%)

Estimators	Max Depth of Weak Learners	Learning Rate
50	1	0.1
100	3	**0.5**
-	**5**	1

Results

- Confusion Matrix created for each model and common evaluation metrics were computed
  - Important to think about what each metric means for the dataset
  - False positives and false negatives both have detrimental real-world impacts (F1 Score most meaningful)
- Both models significantly outperform the probabilistic guess
- Adaboost model is the overall best performing
  - Still ample room for improvement (1/3 of all seller's markets missed!)



Sources

Housing Data. Zillow. (2023, April 1). Retrieved April 17, 2023, from https://www.zillow.com/research/data/

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Project GitHub Repository

For access to all csv files and a Jupyter Notebook with a detailed description of the entire project and findings please visit my GitHub Repository

https://github.com/arwhit/housing-markets-AdaBoost