README.md

Hurricane Impact on Florida Real Estate Value

Final Project: Analyzed the impact of five different hurricanes on Florida Real Estate value using data sourced from Zillow and The National Oceanic and Atmospheric Administration (NOAA).

Summary

For my capstone project at Flatiron School I used machine learning and classification models to analyze hurricane impact on real estate value. Data from NOAA and Zillow were used to provide the features for the model. The target value was whether or not homes in a region increased in value (encoded as a boolean). If the percentage change in the value of a home was in the 75th percentile or greater six months after a hurricane it was considered to be in the category of increase. The model features were average wind speed (AWND), fastest 2 minute wind gust (WSF2), size rank of the city (SizeRank), and home value six months before the hurricane (before). The features AWND and WSF2 were used to assess how severely a city had been impacted by a hurricane.

Objectives

My main objective for this project was to help real estate companies understand how hurricanes impact real estate value in an area. Using machine learning classification algorithms I wanted to know if hurricanes hitting an area could be a predictor of home value increase. As a Floridian and survivor of hurricane Charley (2004) and hurricane Ian (2022) I have directly seen the impact hurricanes have on real estate and wanted to see if that could be modeled. The first and most direct impact is the destruction of homes and properties. Followed by the migration of people out of Florida who proceed to sell their homes, either as a result of increasing insurance prices or the trauma of surviving a hurricane. And while having to rebuild our communities is an arduous task–the silver lining is that properties and infrastructure are improved and updated, increasing both home and property value.

Data

Data Gathering

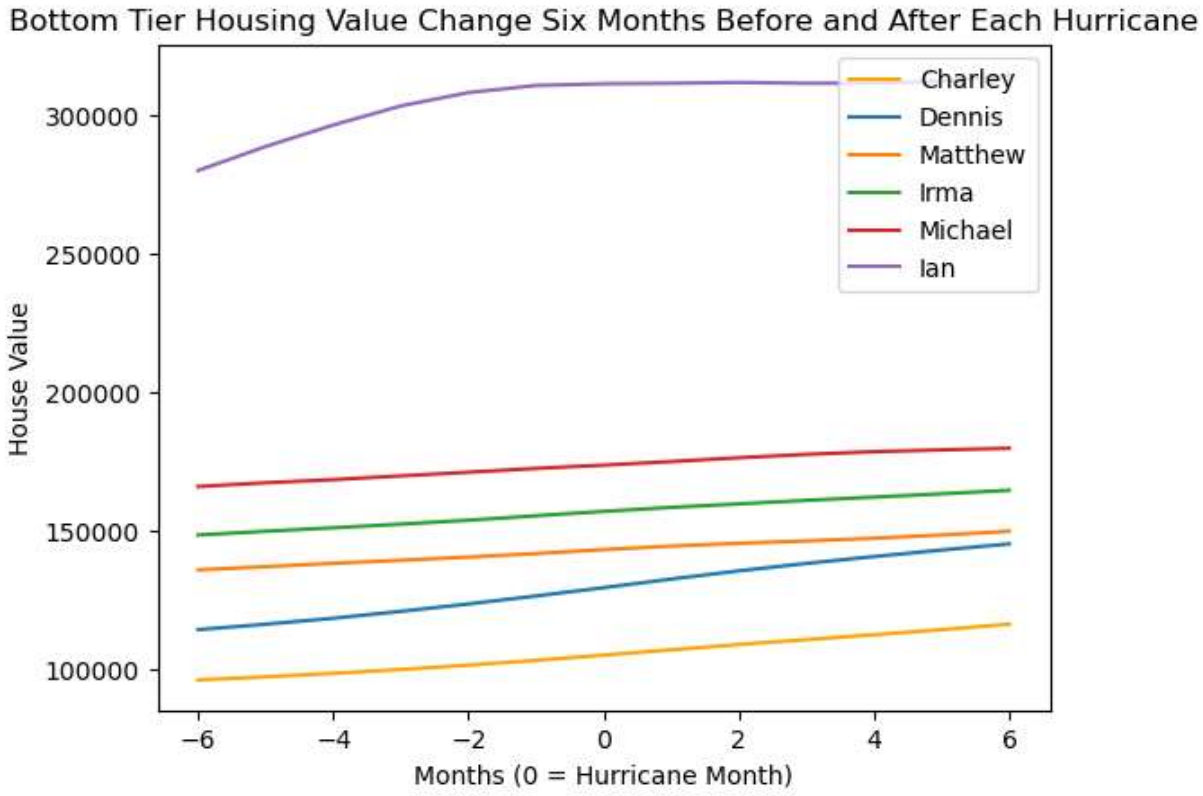
In total I used 9 datasets sourced from Zillow and NOAA. Data from Zillow was used in order to provide home value, city, size rank, and home value six months before and six months after the hurricanes. Data from NOAA was used to know the average wind speed and fastest 2 minute wind gust cities in Florida experienced during six hurricanes. Data from both sources was joined into a final dataset and used for the classification models.

Zillow

Data from Zillow's Home Value Index (ZHVI) was used. Zillow Home Value Index is a measure of a home's typical value and market changes across a given region and type of housing. By measuring monthly changes in property across different housing types and geographies Zillow is able to capture how the market price changes and not just the changes in the kinds of markets or property types that sell on a month to month basis. The ZHVI dollar amount is representative of the "typical home value for a region" and not the "median home value".

I looked at three different datasets from Zillow:

Bottom Tier Homes: typical value for homes within the 5th to 35th percentile range for a given region.
Middle Tier Homes: typical value for homes within the 35th to 65th percentile range for a given region.
Top Tier Homes: typical value for homes within the 65th to 95th percentile range for a given region.

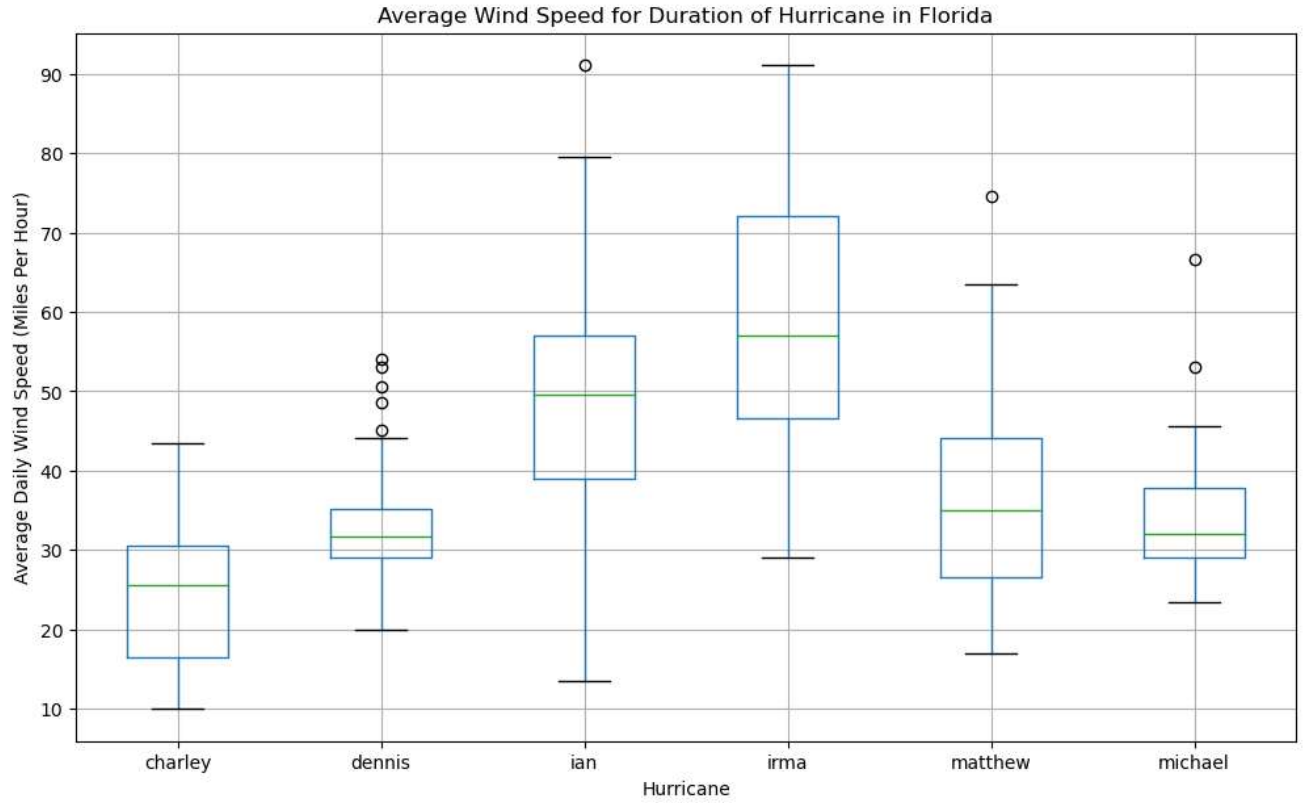


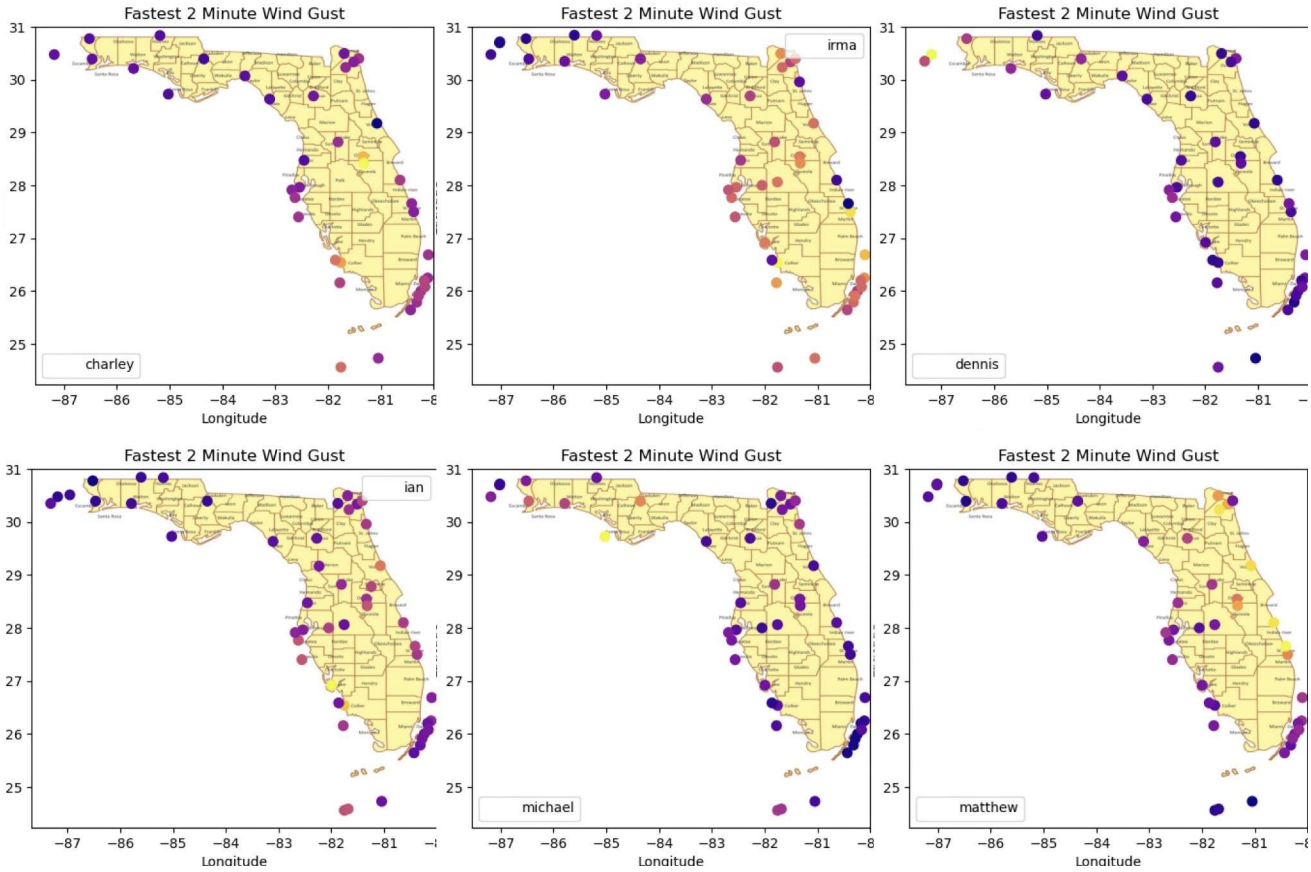
NOAA

Data was obtained from the National Oceanic and Atmospheric Administration (NOAA) and National Climatic Data Center (NCDC) using the Climate Data Online (CDO) database. The CDO provides free access to NCDC's archive of global historical weather and climate data in addition to station history information. These data include quality controlled daily, monthly, seasonal, and yearly measurements of temperature, precipitation, wind, and degree days as well as radar data and 30-year Climate Normals. For the purpose of this project all data was downloaded in CSV format, but PDF and Daily Text was available as well.

Data from NOAA was selected because it provides daily summaries for average wind speed and fastest 2 minute wind gusts for the six hurricanes I wanted to examine.

For the purpose of this project I looked at six hurricanes: Charley (08/2004), Dennis (07/2005), Matthew (10/2016), Irma (09/2017), Michael (10/2018), and Ian (09/2022). I used information from hurricane Charley, Dennis, Matthew, Irma, and Michael to create classification models and validated the model with recent data from hurricane Ian.

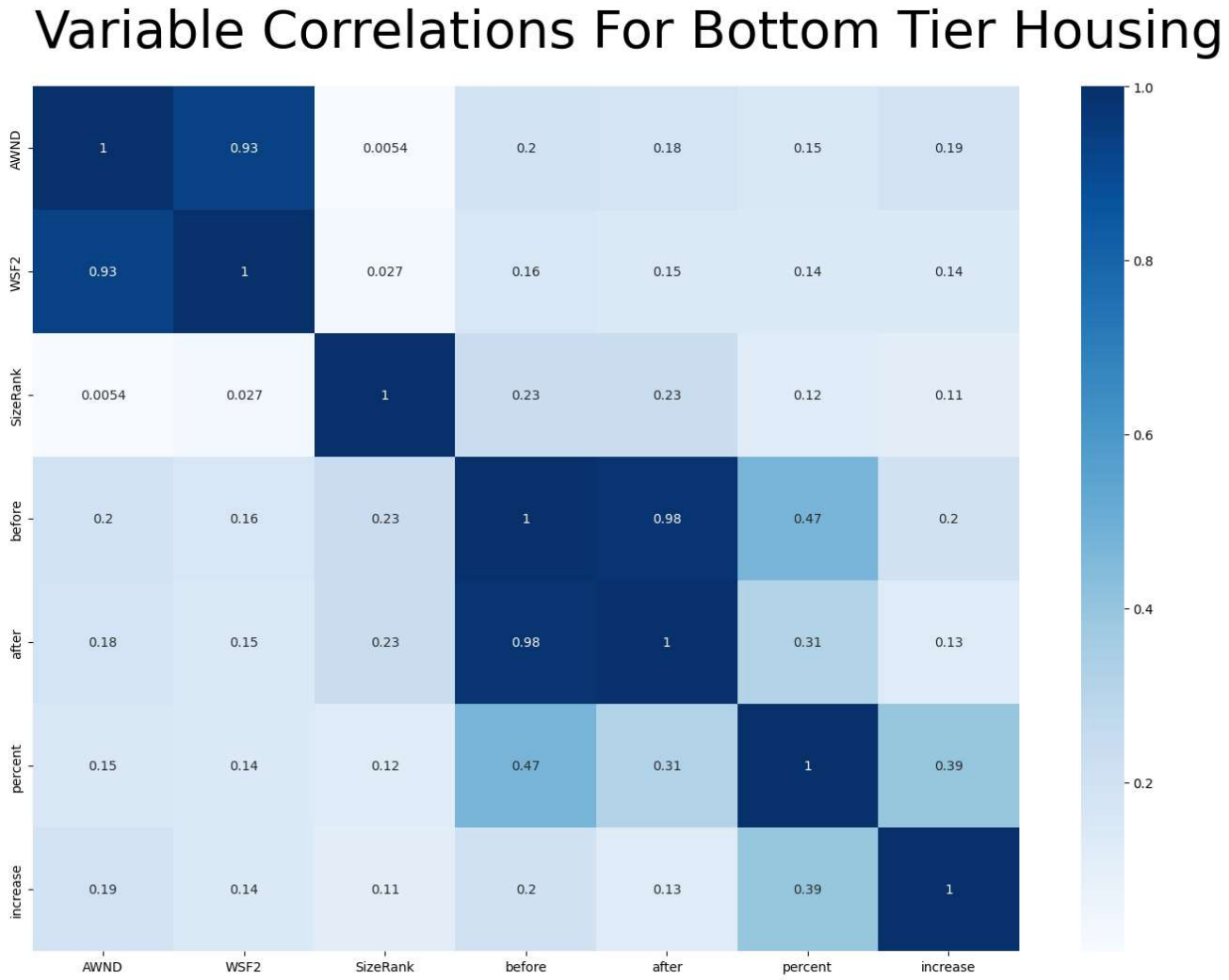




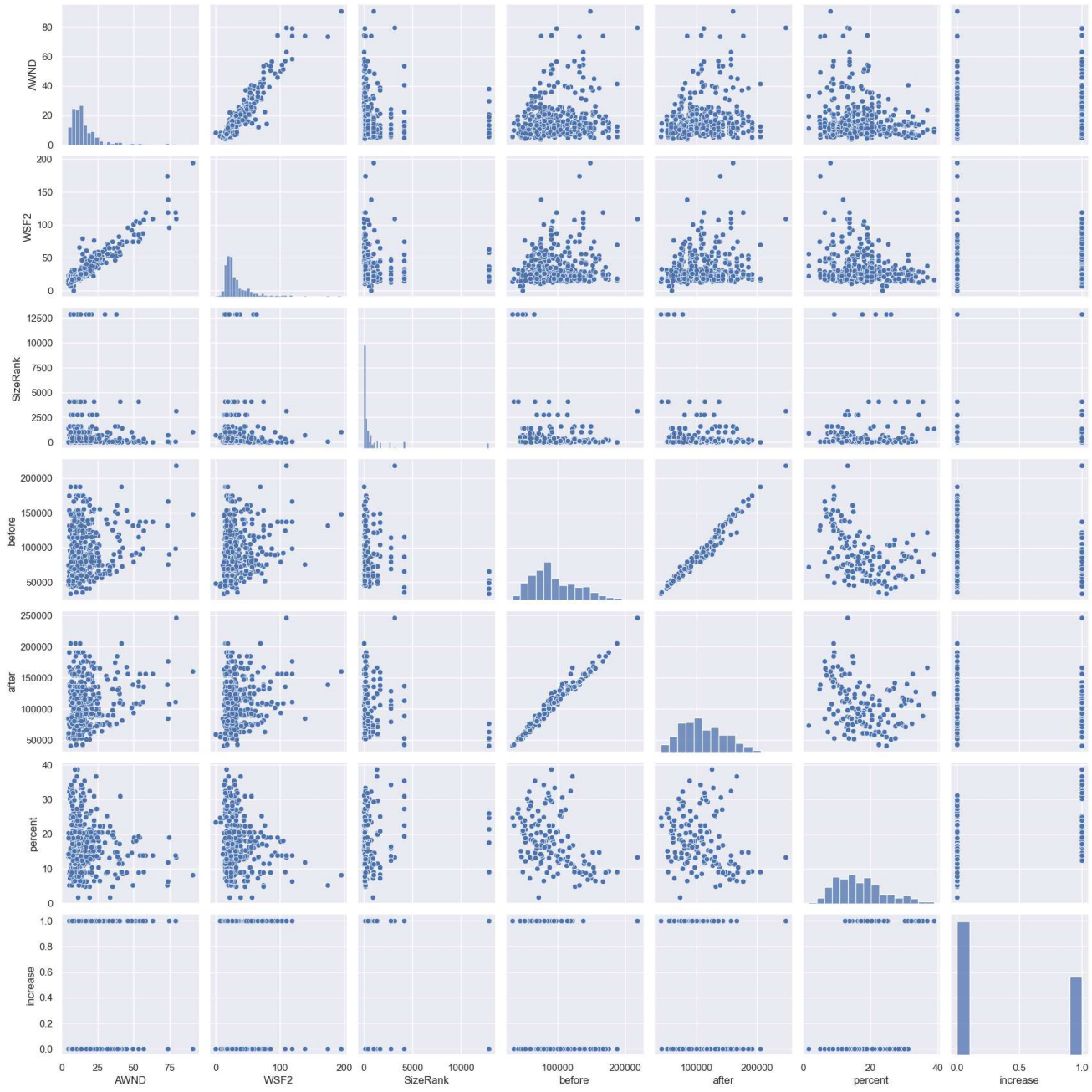
Final Dataset

City: City name
HurricaneName: Name of the hurricane
AWND: Average daily wind speed (miles per hour)
WSF2: Fastest 2-minute wind speed (miles per hour)
SizeRank: Numerical rank of size of cities, ranked 0 through 30,132
before: Home value six months before the hurricane
after: Home value six months after the hurricane
percent: Percent change in home value from six months before hurricane to six months after hurricane
increase: 1 = increase of 75% of more in home value, 0 = no increase of 75% of more in home value

I trained the models on data drawn from the bottom tier home value dataset which had 141 entries. I chose this dataset to train our models on because it had the best target variable class imbalance of 68%. I used the dataset containing all home values to assess the impact of wind features and tune the model which contained 344 entries. I also used a dataset containing just data from hurricane Ian to see how the model would perform on future hurricane data. The hurricane Ian dataset contained 27 entries.


Variable Correlations For Bottom Tier Housing
AWND
WSF2
SizeRank
before
after
percent
increase
1
0.93
0.0054
0.2
0.18
0.15
0.19
0.93
1
0.027
0.16
0.15
0.14
0.14
0.0054
0.027
1
0.23
0.23
0.12
0.11
0.2
0.16
0.23
1
0.98
0.47
0.2
0.18
0.15
0.23
0.98
1
0.31
0.13
0.15
0.14
0.12
0.47
0.31
1
0.39
0.19
0.14
0.11
0.2
0.13
0.39
1
1.0
0.8
0.6
0.4
0.2

Bottom Tier Housing Pairplot



Feature Engineering

In order to create our models I needed to engineer two additional features: home value increase and city.

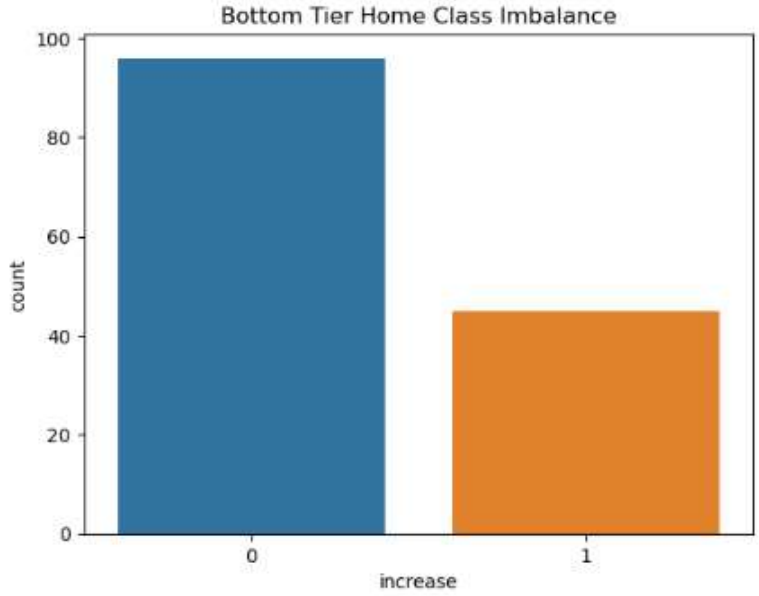
Using GeoPy to Get Cities

Data from NOAA provided location information in the form of latitude and longitude while data from Zillow provided location information in the form of city. In order to join the data from NOAA to the data from Zillow I needed to know the city names. Using the coordinates provided by the NOAA dataset I used GeoPy to reverse geolocate the city names.

Home Value Increase

Since this was a classification problem the target variable had to be in the form of a class. Using home value data from six months before and after each hurricane I engineered a column that contained two categories, increased significantly (coded as 1) and did not increase significantly (coded as 0). If the percentage change in the value of a home was in the 75th percentile six months after a hurricane it was considered to be in the category 1, if not then 0.

```
0    96
1    45
Name: increase, dtype: int64
majority percent is 68.08510638297872 %
```



Modeling

Actuals	Home Value Did Not Increase More Than 75%	True Negative (TN): Home value predicted not to increase and home value did not increase	False Positive (FP): Home value predicted to increase and home value did not increase
	Home Value Increased More Than 75%	False Negative (FN): Home value predicted not to increase, but home value did increase	True Positive (TP): Home value predicted to increase and home value did increase
Hurricane Impact on Real Estate in Florida Confusion Matrix		Home Value Did Not Increase More Than 75%	Home Value Increased More Than 75%
		Predictions	

Of equal concern to the business problem was the model predicting False Negatives or False Positives. In this situation both are of equal importance to our real estate client. If the model predicts that the home value will not increase but it does (false negative) then our client could lose out on a possible higher return on a home. However, if the model predicts that the home value will increase but it does not (false positive) then our client may have invested money into a home that does not pay off.

Metrics

The metrics used were accuracy and F1 Score to assess model performance. Since there was a class imbalance I could not solely rely on accuracy to communicate model performance. Since, precision and recall were of equal importance for our business problem F1 score was used to assess model performance. The model's AUC-ROC curve was also assessed to see how well the model performed.

Accuracy

Model accuracy is a machine learning classification model performance metric that takes the ratio of true positives and true negatives to all positives and negative results. It communicates how often our model will correctly predict an outcome out of the total number of predictions made. However, accuracy metrics are not always reliable for imbalance datasets. Accuracy Score = (TP + TN)/ (TP + FN + TN + FP)

F1 Score

Model F1 Score is a model performance metric that gives equal weight to both the Precision and Recall for measuring performance. F1 Score = 2* Precision Score * Recall Score/ (Precision Score + Recall Score/)

ROC

The Receiver Operator Characteristic (ROC) curve is used to assess a model's ability to correctly classify by plotting the true positive rate against the false positive rate. A curve that 'peaks' more quickly communicates that there is a good true positive rate and a low false positive rate. The area under the curve (AUC) is derived from ROC and has a baseline chance of 50% accuracy, hence, an AUC closer to 1 signifies a better classification model.

Baseline

As a baseline I looked at the class imbalance of the target variable for the various datasets. This allowed me to know if the model accuracy was any better than just selecting the majority class.

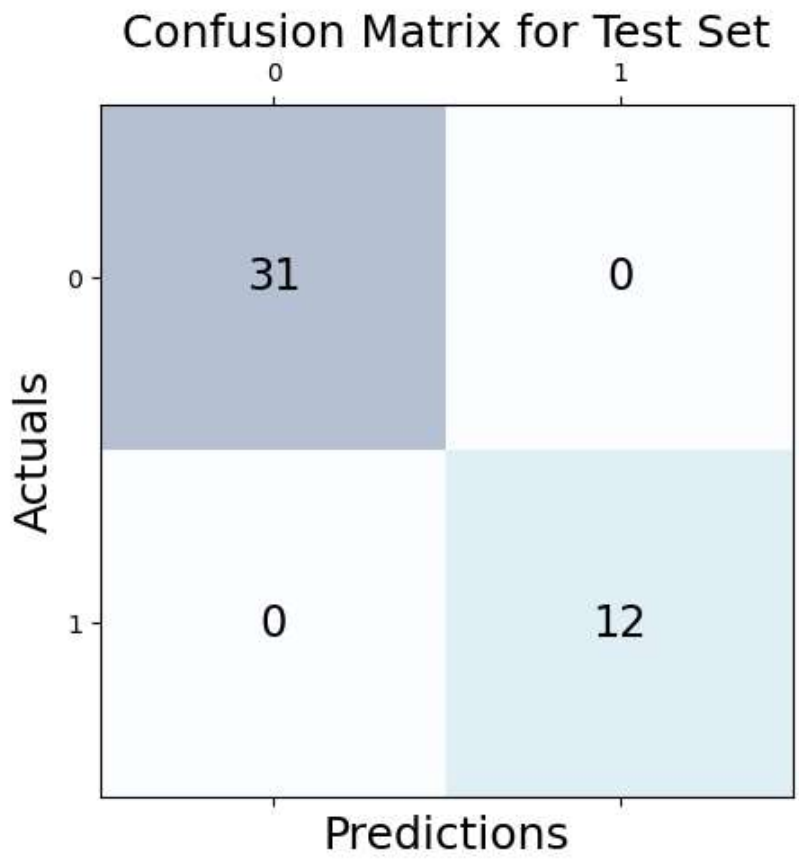
Dataset	Increase	No Increase	No Increase as Percent
Bottom Tier	45	98	68%
Middle Tier	33	106	76%
Top Tier	28	114	80%
All	89	255	74%

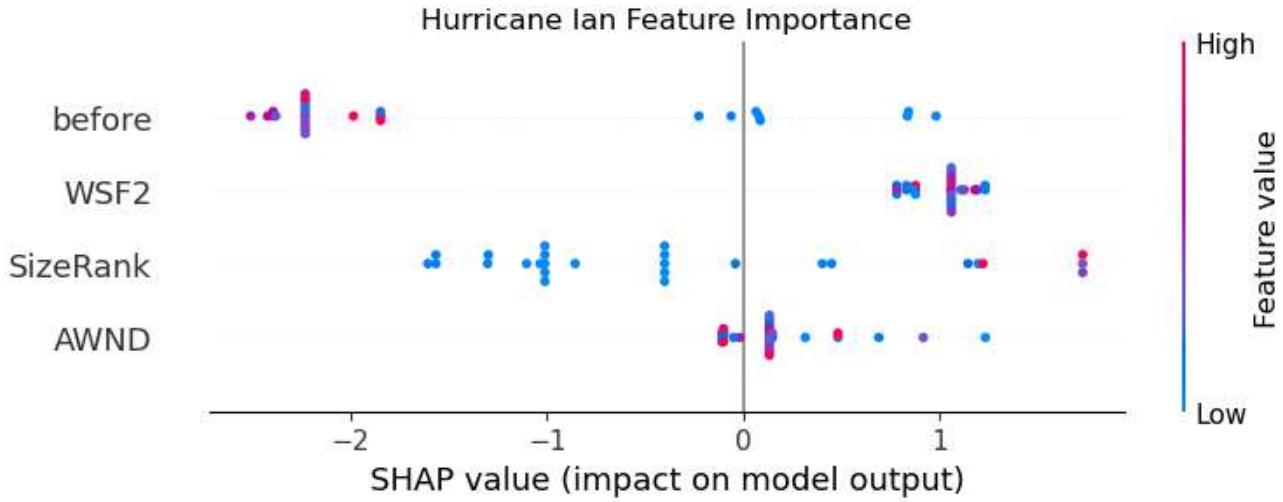
Logistic Regression

I chose to use a logistic regression model because it is commonly used for prediction and classification problems. The model by default was set with an L2 penalty. It was iterated through the scaled data. Iterations were also performed after removing collinear variables and using SMOTE to adjust for the class imbalance. The logistic regression model that performed the best was the SMOTE model with no collinear features and had an accuracy of 0.74 which was slightly better than our baseline accuracy of 0.68 and an F1 score of 0.68. It found that AWND was the most important feature in predicting home value increase.

XG Boost

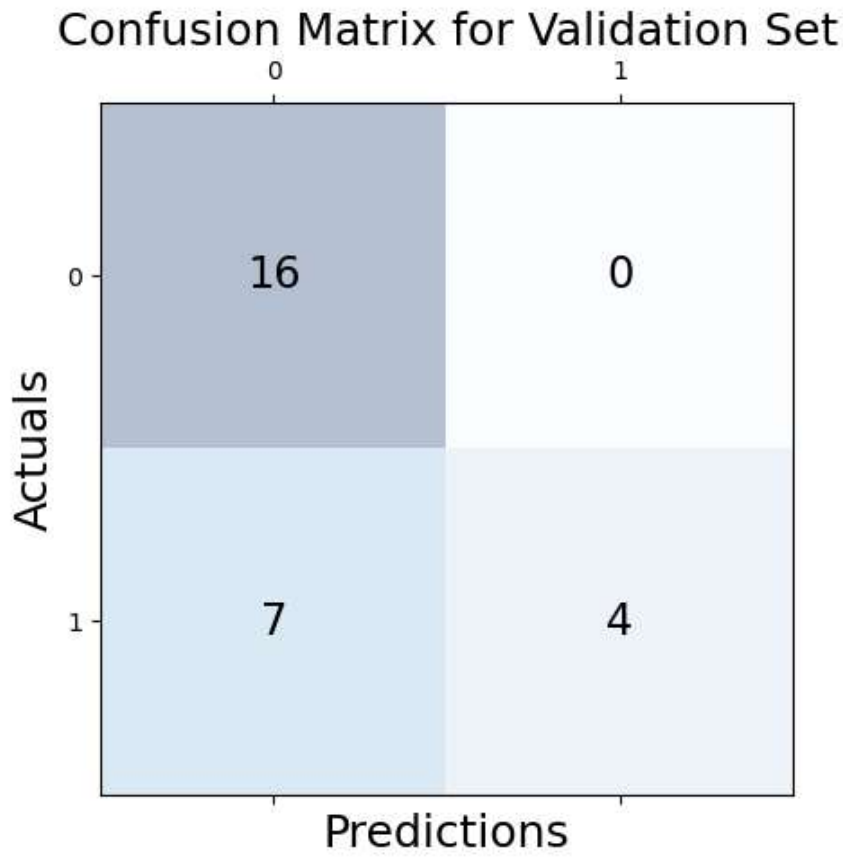
An XG Boost model was used due to its adaptability and strong prediction performance. Our XG Boost model had perfect accuracy and a perfect F1 score. When the model was run without the inclusion of wind features model accuracy dropped to 98% and the F1 score dropped to 0.96. Model tuning was also performed on a model that contained just wind features, however, it did not improve overall model performance.





Final Model Validation

Our final model was an XG Boost model. Since the goal for this project is to help real estate agencies understand how hurricanes impact home value data from a recent hurricane was used to test our model. Using data from hurricane Ian our model performed with 74% accuracy, which is slightly better than the baseline accuracy of 59%. The F1-score was 0.544 and the model predicted 7 false negatives and no false positives. The AUC score was 0.92.



Recommendations

I recommend further looking into cities that have large Size Ranks and lower than average housing prices as a predictor, regardless of if a hurricane happened because it is a strong indicator of home value. Hurricane features do slightly improve model performance and should also be considered.

Wind speed features have a correlation that is similar to that of SizeRank and before prices, hence, it may be an indicator of home value increase and it is relatively relevant to models predictions.

I recommend using the model to attempt to buy homes after the next hurricane to see if the model is effective in making business decisions.

Further Improvements

Further improvements for this project would be to continue collecting data. For the scope of this project I looked at all hurricanes that were a category 4 or above for the last twenty years. I recommend continuing at all hurricane data that also has real estate data available.

I also recommend looking at a different region which is impacted by hurricanes, such as Texas, to see what other features could be of importance while modeling.

Repo Structure

Structure

|-- data <- Folder containing datasets used in analysis, images, and documentation
|-- presentation and notebook pdf <- Folder containing presentation and notebooks pdf
|-- .gitignore <- File that ignores files
|-- 0.PresentationNotebook.ipynb <-Narrative documentation of analysis in Jupyter notebook |-- 1.HousingData.ipynb <- Narrative documentation of analysis in Jupyter notebook
|-- 2.HurricaneData.ipynb <- Narrative documentation of analysis in Jupyter notebook
|-- 3.JoiningDatasets.ipynb <- Narrative documentation of analysis in Jupyter notebook
|-- 4.Modeling.ipynb <- Narrative documentation of analysis in Jupyter notebook
|-- 5.WindSpeedAnalysisandModelTuning.ipynb <- Narrative documentation of analysis in Jupyter notebook
|-- README.md <- The top-level README for reviewers of this project
|-- References.txt <- List of works cited
|-- ReproductionInstructions.txt <- How to reproduce this project
└ - environment.yml <- Environment used in this project

Contact Me

Thank you for checking out my GitHub if you have any further questions please contact me at samgrasland@gmail.com!

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%