dukebaya Update README.md 02b0f0c · now History

88 lines (57 loc) · 8.06 KB

Preview Code Blame Raw

Identifying High-Potential Real Estate Inventment Opportunities:

Apex Assets Investment is an investment firm that buys and sells real estate in the USA and is facing challenges as to how they should make their investments. They have consulted us to provide guidance this issue to ease their process. We have used data from .Zillow Research. to achieve this objective. Zillow Research publishes top-tier Zillow Home Value Index (ZHVI): A measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range.We have cleaned the data, forward filling of missing values, and anything else that would disturb the research. Time Series Modelling has been used in this notebook to produce future forecast property values based on previous observations, to identify underlying trends/ patterns within the data over time and to detect any signs of seasonality within the data provided. These factors will contribute towards determining how Apex Assets Investment build their portfolio.

2. Business Understanding

2.1. Overview

Apex Assets Investments is a real estate firm which provides state of art opportunities in the real estate scope such as housing. However, it is grappling with a cahllenge on the beest avenues to do the investing. The primary goal of this project is to to analyse the factors that will determine the best zipcodes to invest in and build and time series model that will forecast the real estate prices of various zipcodes and provides insights and recommendations based on the built time series model in this project. The target stakeholders are executives, product managers, marketing teams, finance and account teams, and contact centre team at Apex Assets Investments. The data used can be accessed from here.

2.2. Business Problem

In today's competitive real estate market, Apex Assets Investments faces the challenge of finding the best lucrative places to invest. With countless zip codes, fluctuating prices, the pursuit of optimal investment destinations poses a the task that can feel overwhelming. The objective is clear: pinpoint the top 5 zip codes with the highest potential for profit while managing risks wisely.

To tackle this challenge, Apex Assets Investments turns to data and analysis using the time serries model. By studying past trends and using advanced forecasting techniques, they aim to uncover hidden opportunities in different zipcodes. But it's not just about numbers—it's also about understanding communities, local developments, and what makes each area unique.

With this approach, Apex Assets Investments can make informed decisions that not only benefit their investors but also contribute positively to the neighborhoods they invest in. Ultimately, it's about finding the right balance between growth and stability in the ever-changing world of real estate.

3. The Dataset Understanding and Preparation

The dataset used contains housing information from July 1996 to April 2018 obtained from here.The dataset has numerous features such as RegionID, RegionName, City, State, CountyName etc. distributed across 14723 rows and 272 columns. Of these 272 columns, 7 of the columns are described below. The remaining 265 columns are records of the given Zip code's average home value between April, 1996 and April, 2018. The other 7 columns are:

'RegionID'- Each record contained a unique ID number. *'RegionName'-This is the column that provides the Zip code for the given row. *'City'- The name of the city in which the Zip code is located. *'State'- The name of the State in which the City is located. *'Metro'-The name of the Metro region. *'CountyName'-The name of the County *'SizeRank'- The ranking of the particular Zip code's size relative to other records in the dataset.

Our project goal was defined by our Business Problem where our stakeholder was only interested in purchasing a home within a zipcode with high Return On Investment.

3.1. Data Preparation

We investigated the data to identify:

- Previewed the dataset.
- Checked the dataset , analytics for missing values, datatypes and summary
- Checked the unique values in the first 7 columns

3.2. Feature Engineering

We did some conversion of the data into the below formats:

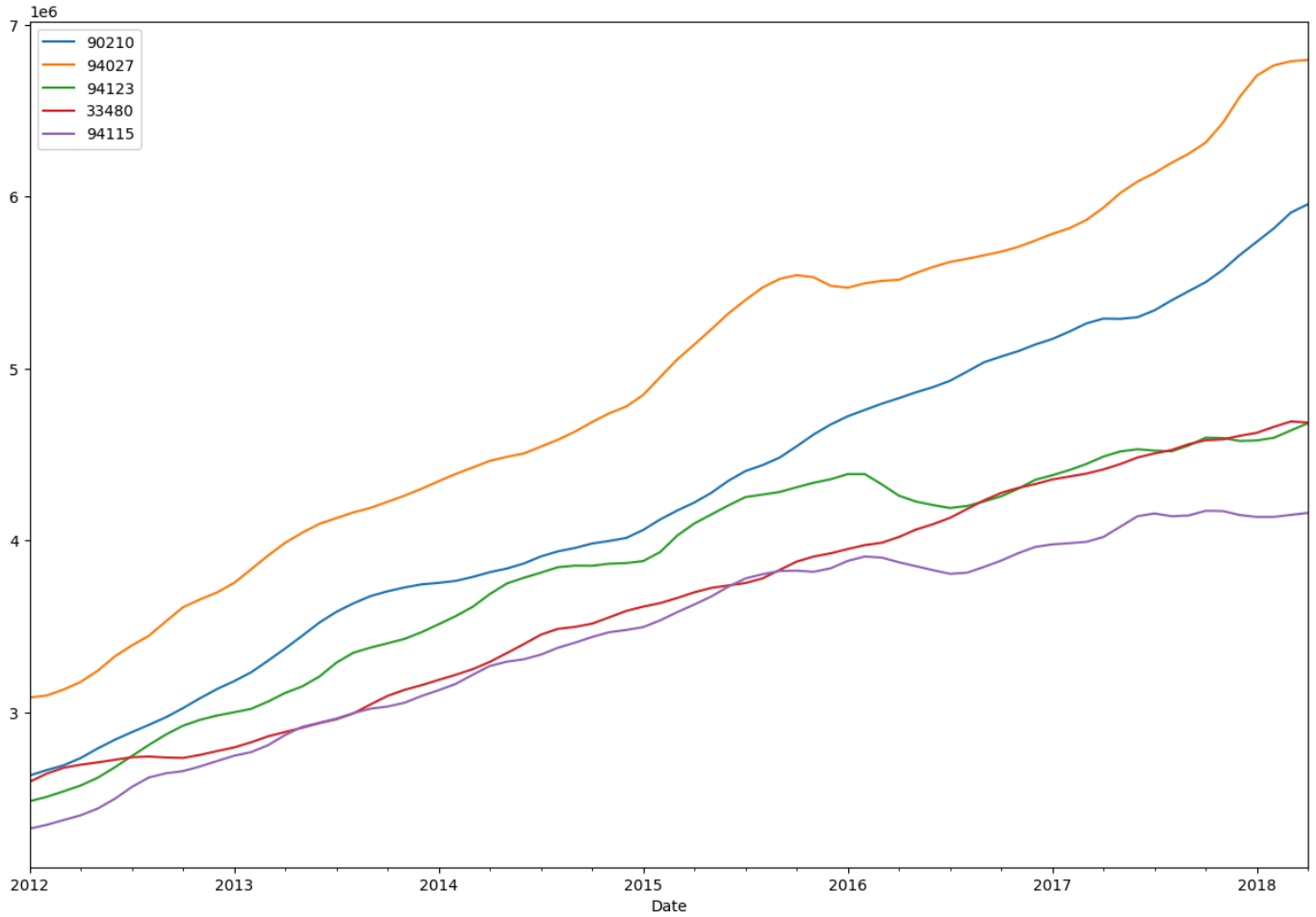
- Change the dates to datetime values
- Rename the RegionName to Zipcodes

3.3. EDA

Having cleaned our dataset to only include Zip codes with top 10 prices we were left with the top 10 zipcodes displayed below. The 10 zipcodes seen here consisted of unique ID Zip codes between them. We were also able to explore the below:

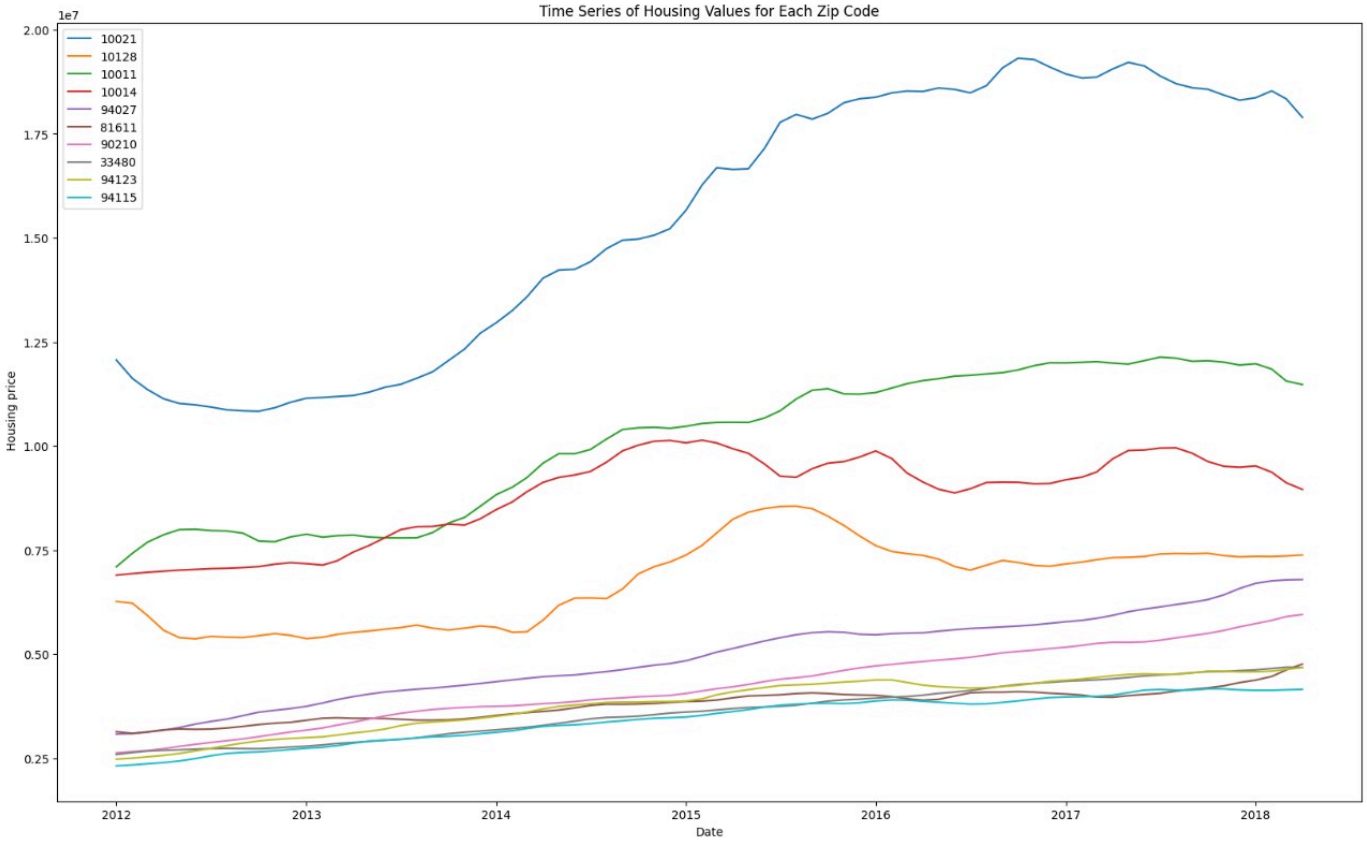
- Summary Statistics

- Top 5 zipcodes with high ROI



)

- Analysis of the top 10 zipcodes Value, ROI,CV



)

4. Modeling

Train Test Split: The script splits the time series data into training and testing sets, ensuring 80% of historical data is allocated for training to train the model effectively while keeping 20% for evaluation. The Performance baseline before modeling gives a ROI of 21.39% from 2016 to 2018. This will later be compared to the modeling and prediction results

To find the five zipcode areas that are the most optimal for real estate investment , We ran an Auto Arima time series models on every all the zipcode areas of concern individually by means of functions .

We used this model due to the nature of data and it various advantages such as the ability automatically selects the optimal values for the p, d, and q parameters, as well as seasonal parameters (P, D, Q, m) if applicable. This removes the need for manual parameter tuning, saving time and effort.

	Actual	Predicted	% Error
Mc	1,253,800	1,251,722	0.17
Sea Is	2,440,000	2,431,463	0.35
Da	1,388,100	1,383,292	0.35
San Franc	3,345,857	3,330,054	0.47
Berk	1,974,100	1,962,974	0.56
...	...	...	...
Menlo	2,561,200	1,925,987	24.80
Brook	2,000,850	1,499,232	25.07
Glenb	2,161,900	2,715,748	25.62
Sullivans Is	1,837,000	2,475,610	34.76
Key Bisc	2,871,700	3,880,653	35.13

82 rows × 3 columns

5. Evaluation

The script evaluates the accuracy of the trained models by comparing their predictions on the testing set with actual values, providing insights into the performance of the models in predicting future trends. On average the model's predictions were roughly 8.2% off from the actual values from our test set. The model is giving a lower ROI. Since we have a low margin of error we are comfortable to proceed with the mordeling.

6. Conclusion

In our investigation to assess the effectiveness of our models, we uncovered compelling evidence suggesting their efficacy. By employing an exploratory data analysis (EDA) approach on our training dataset to identify five cities for investment and subsequently simulating investments in those selected cities on our test dataset, we realized a notable 21.39% return on investment. Comparatively, when leveraging modeling techniques to predict the optimal five cities for investment within New York over the duration of our test dataset, we attained a respectable 17.27% return on investment. This analysis underscores the value of our models' recommendations, indicating that even if their predictive accuracy is not exceptionally high, their insights remain valuable..

7. Recommendations

These are the top 5 zipcodes by ROI one year out, as predicted by our models, and serve as our final recommendations.

	City	Current Value	Predicted Value	Net Profit	ROI
1	Al	3,069,100	3,128,438	59,338	1%
2	Amagan	3,141,100	3,369,837	228,737	7%
3	A	4,766,600	5,442,671	676,071	14%
4	Athe	6,796,500	6,898,057	101,557	1%
5	Av	1,665,600	1,685,437	19,837	1%
6	Berk	1,974,100	2,002,863	28,763	1%
7	Beverly H	3,899,300	3,903,336	4,036	0%
8	Boca Gr	1,989,100	2,077,214	88,114	4%
9	Bridgeham	2,592,100	2,895,226	303,126	11%
10	Brook	2,000,850	2,125,373	124,523	6%
11	Burlin	2,964,000	2,978,624	14,624	0%
12	Cambr	2,037,600	2,171,339	133,739	6%
13	Carmel-by-the	1,603,550	1,624,524	20,974	1%
14	Chil	1,526,200	1,546,644	20,444	1%
15	Cold Spring Ha	1,128,900	1,135,776	6,876	0%
16	Coro	1,997,600	2,024,888	27,288	1%
17	Cuper	2,490,200	2,330,157	-160,043	-6%
18	Da	1,388,100	1,391,374	3,274	0%
19		1,697,200	1,725,024	27,824	1%
20	Del	2,139,100	2,141,834	2,734	0%

Aspen Bridgehampton Amagansett Brookline Cambridge

For More Information

For more info please review our full analysis in notebook and slides

Authors

1. Caroline Njoroge
2. Miriam Ongare
3. Mercy Ronoh
4. Philip Mweri
5. Chepkemoi Chepkemoi