**Housing Price Prediction in Colorado**

Regis University

MSDS692 – Data Science Practicum 1

Professor: Christy Pearson

By

Sangita Lamichhane

2022

**Project Overview**

The real estate industry comprises of varieties of stakeholders ranging from regulatory bodies, private sectors and investing firms deeming it to be one of the vital sectors for revenue generation. These stakeholders have a high demand for better understanding of the real state’s growth mechanism and its return of investments (ROI). Colorado offers great opportunities for investment in housing market as recently it is observed that several real estate investors are interested to invest in this state. This project can be considered as an evidence-backed step towards a rational decision making for the benefit of those investors. This project focused on predicting the most favorable cities in Colorado to invest in and get highest return of investment (ROI) based on data from Zillow. In addition to this, we have performed exploratory data analysis (EDA) on data collected from Redfin to analyze impacting features and find out their correlation to housing prices. This project provides a recommendation on which five Colorado State metro areas are the most prime for investing in now and selling a year from now.

**Research Problem**

The goal of this research project is to explore the model to predict, with time series analysis, which cities in Colorado state are best for real estate investment. This analysis will appeal investors as it will help them make choices by predicting the estimated return of investment (ROI) for their investment in certain cities in the state of Colorado. This study finds out which 5 cities may yield high ROI percentage if invested.

**Data Description**

This project uses the first dataset from Zillow research data to determine the top 5 metro areas for real estate investment in Colorado and forecasts their growth over the next year. The dataset contains the median home prices per metro areas recorded monthly over the period from April 1996 to August 2022. The other dataset used for exploring the relationship to housing price is from Redfin.

The dataset “Metro\_zhvi\_uc\_sfrcondo\_tier\_0.33\_0.67\_month.csv” was downloaded from Zillow where we have 899 rows, with 325 columns. Of those 325 columns, 5 of them describe the metro area represented by each row. Columns 6 through 325 represent a time series that spans from January of 1996 to August of 2022. The values are the median home price for that metro area on that specific month. The dataset has 899 rows, the first column is the only row that represents a region other than a metro area which tracks the median home price for the entire United States.

The dataset “county\_market\_tracker.tsv000” was downloaded from Redfin where we have 699630 rows and 58 columns. This dataset is used for exploration of finding out correlation with other factors during the same period of our analysis.

**Methodology**

In this project we calculated the return of investment (ROI) for metro areas of Colorado through an EDA approach as well as ARIMA model fitting. Since, each of the 17 cities required their own model, which had to be grid searched, fit, and used to generate predictions, we have used a custom function to streamline the process. These functions ran auto Arima, a grid searching function, fit an ARIMA model with the optimal parameters, and made predictions using the fit models. Also, to verify and compare the results from the ARIMA model, we also carried out analysis through Facebook Prophet.

**EDA approach**

Chart, line chart

Description automatically generated

Figure Average median home price in CO State from 1996-2021

Chart, treemap chart

Description automatically generated

Figure Heatmap showing correlation of factors present in Redfin Data

Off market in two weeks and sold above price shows positive correlation. The more the houses are off market in two weeks the more the selling price of the houses. During the same time, it is observed that new listings are sold above the listing price as well as no significant price drop were observed as shown in the above heat map. In addition to this, its observed that the median list price of the houses does not make any difference to the number of houses sold showing no significant correlation between the median list price and the number the house sold.

**Chart, bar chart

Description automatically generated**

Figure Top five choice based on EDA approach

Table

Description automatically generated with medium confidence

Figure ROI of top five choice based on EDA approach

Based on EDA approach, it is discovered that the top five metro areas with highest ROI are Denver, Greely, Fort Morgan, Boulder, and Pueblo. We have graphed the sales price distribution for the entire state and to see the overall trend we have plotted time series average median home prices in Colorado state as shown in Fig-1. Using an EDA approach, we could have achieved 36.49% ROI from 2020 to 2022.

**ARIMA Model**

Table

Description automatically generated

Figure Accuracy of the ARIMA model

Using ARIMA model, for accuracy we compared and validated the predicted value with the actual data from 2022.

Table

Description automatically generated

Figure Prediction of ARIMA model for August of 2022

From Fig 5, Pueblo, Colorado Springs, Grand Junction, Sterling, and Fort Morgan are the top 5 metro area from ARIMA Model for trained data.

Table

Description automatically generated

Figure Prediction of ARIMA model for August 2023

From Fig-6, we identified that the most ROI for 2023 can be achieved from investment in Grand Junction, Colorado Springs, Edwards, Durango, and Breckenridge. Using ARIMA modeling, we could have achieved 38.51% ROI from 2020 to 2022

**Facebook Prophet**

Graphical user interface, table

Description automatically generated with medium confidence

Figure Prediction of Facebook Prophet model for August 2023

Using Facebook Prophet model, we observed that Pueblo, Breckenridge, Durango, Colorado Springs, and Steamboat Springs are identified to provide highest ROI.

**Findings and Conclusion**

**Chart, line chart

Description automatically generated**

Figure One year forecast for Grand Junction, CO

**Chart, line chart

Description automatically generated**

Figure One year forecast for Edwards CO

**Chart, line chart

Description automatically generated**

Figure One year forecast for Colorado Springs CO

**Chart, line chart

Description automatically generated**

Figure One year forecast for Breckenridge CO

**Chart, line chart

Description automatically generated**

Figure One year forecast for Durango CO

In our analysis, we can see that ARIMA models trained on data from 2012 - 2020 and tested on data from 2020- 2021 fairly outperformed an EDA based approach on the same time periods which shows that applying modeling was worth our effort. During simple EDA based analysis for selecting which five cities to invest in resulted in an ROI of 36.49% where the ARIMA model-based approach resulted in an ROI of 38.51%. Furthermore, it should be taken into consideration that the ARIMA based analysis had an error of 19.06% i.e., the predicated value deviated. 19.06% from the actual values for the test period. Apart from this, for the comparison of the model’s prediction we have also performed Facebook prophet model prediction which gave us top five cities which included the same three cities from our ARIMA model showing us that the model performed well. The graphs shown above provides us the predictions of top five cities with highest predicted ROI.

**Future Consideration**

There are other factors that I would also like to study about these metro areas to better understand whether median home price can really keep growing as predicted. For instance, the information about wage growth in Colorado, migration data as well as inflation data to determine whether prices in the Colorado metro areas are already close to the maximum the market will allow. Also, the ROI percentage obtained seems high in some metro areas, it would be better to do more research to figure out the sustainability of the rapid growth of house prices seen in recent years. Further analysis may include building linear model with additional features such as federal income tax rate and housing market safety score and more.

# Reference

ademos. (2022, August). *Chapter 23: Using ARIMA for Time Series Analysis.* Retrieved from ademos: https://ademos.people.uic.edu/Chapter23.html

Agent, A. H. (2022, Oct). *American Homes Agent.* Retrieved from Is the Colorado Housing Market Slowing Down?: https://americanhomeagents.com/home-trends/is-the-colorado-housing-market-slowing-down

otexts. (2022, August). *Prophet model.* Retrieved from otexts: https://otexts.com/fpp3/prophet.html#ref-prophet

Redfin. (2022, August). *Redfin.* Retrieved from Data Center: https://www.redfin.com/news/data-center/

Stack, A. (2021, November). *Zillow\_Real\_Estate\_Time\_Series\_analysis.* Retrieved from GitHub: https://github.com/AidanStack/Zillow\_Real\_Estate\_Time\_Series\_analysis

Vidhya, A. (2021, November 12). *A Comprehensive Guide to Time Series Analysis.* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/

Zillow. (2022, August). *Housing Data.* Retrieved from Zillow: https://www.zillow.com/research/data/