Data Analytics Capstone Topic Approval Form

Student Name: Ubaldo Martinez III

Student ID: 000322245

Capstone Project Name: Time Series Forecasting of United States Home Sales

Project Topic: This endeavor aims to leverage data provided by Zillow, encompassing residential property transactions spanning the years 2008 to 2018. The primary objective is to construct a robust time series forecasting model adept at predicting home sales, encompassing the entirety of 2019 and projecting into the subsequent years until 2022.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: Can home sales in the United States be effectively forecasted based solely on research data?

Hypothesis:

Null hypothesis- A predictive time series forecasting model with a mean absolute percentage error of < 20% cannot be generated from the research dataset.
Alternate Hypothesis- A predictive time series forecasting model with a mean absolute percentage error of < 20% can be generated from the research dataset.

Context: Zillow is a prominent online real estate marketplace that provides users with comprehensive information about properties, rentals, and home values across the United States. The platform offers a wide range of data related to real estate, including property listings, historical sales data, property values, and rental information. This data can be utilized in data analysis projects to gain insights into the real estate market, track property trends, assess property values over time, and make informed decisions related to buying, selling, or renting properties. Zillow's extensive dataset and user-friendly interface make it a valuable resource for individuals and professionals engaged in data analysis, real estate research, and market evaluation.

Data: The data needed to attempt to generate a predictive ARIMA/SARIMA model is published by Zillow Research division. The home sales count dataset contains homes sold for the top 94 cities based on population size with exception to the inclusion of Fort Collins, Co (Ranked 150th). The dataset's interval period is in months, spanning from 2008-02-29 to 2023-07-31. Dataset contains the following variables of use:

Field	Data Type
RegionID	Qualitative
SizeRank	Qualitative
RegionName	Qualitative
RegionType	Qualitative
StateName	Qualitative

Dates	Quantitative

A secondary downloadable dataset contains the average home value for single-family, condominiums and co-operative within the 35th to 65th percentile range. The number of cities included in this dataset is 895, ranked by population size. The dataset's interval period is in months, spanning from 1996-02-29 to 2023-07-31. Dataset contains the following variables of use:

Field	Data Type
RegionID	Qualitative
SizeRank	Qualitative
RegionName	Qualitative
RegionType	Qualitative
StateName	Qualitative
Dates	Quantitative

Both Home Sales and Zillow Home Value Index datasets encompass single-family homes, condominiums, and cooperatives defined as follow:

Single-Family Home: a single-dwelling unit, with one owner, no shared walls and on its own land.

> **Condominium:** an individually owned residential unit in a building or complex comprised of other residential units.
>
> **Co-Operative:** a distinctive form of residential housing choice that constitutes a corporation where unit owners lack complete ownership of their respective units. Rather, every inhabitant assumes the role of a shareholder within this corporation, with their stake influenced, in part, by the proportional magnitude of their dwelling.

The central constraint, although not deterring model forecasting accuracy, is the limited number of cities in the Home Sales dataset (94) versus the Average Home Values dataset (895). To maintain data consistency only the 94 cities found in the Home Sales dataset will be utilized instead of all 895.

Data Gathering: Both Home Sales and Home Values datasets are downloadable in CSV format and maintained by Zillow. Home Sales is labeled as "Sales Count Nowcast (Raw, All Homes)" on Zillow's research page, described as an estimated number of unique properties that sold during the month after accounting for the latency between when sales occur and when they are reported. Home Values is labeled as "ZHVI Single-Family Home Time Series ($)" and represents home values for any given region as a weighted average of the middle third of homes.

Home Sales dataset has four regions with one 'NaN' value each and will be resolved by taking the average of the values before and after, to stand-in for this data. One region has two years of missing values and will be removed entirely. The Home Values dataset contains a significant amount of 'NaN' values, most of them pertaining to smaller

regions, therefore will require extensive cleaning. Both datasets are represented in a wide format and will require to be transformed into long format to perform exploratory data analysis.

This analysis primarily focuses on predicting the volume of home sales. While certain aspects of exploratory data analysis will encompass property values, it's important to note that property values themselves are not the target of forecast in this study.

Data Analytics Tools and Techniques: Exploratory Data Analysis will encompass diverse perspectives of both home sales and property values. This will involve dissecting differences across various dimensions, including regional and state-specific home sales and values. This holistic examination aims to glean insights into prevailing trends and seasonal variations within the dataset. Following this comprehensive exploration, the temporal data for nationwide sales count and property values will undergo decomposition to discern the influence of seasonal and trend components on both data series. Subsequently, the dataset will be partitioned into two sets: a training set, encompassing 70% of the observations (spanning 2008 to 2018), and a testing set, incorporating 30% of the observations (spanning 2019 to 2022).

Appropriate time series forecasting model(s) will be developed and tailored to the training data. Subsequently, these models will be utilized to generate forecasts for the designated test set. To evaluate the performance of different models, the mean squared error of the forecasts will be computed and compared. The model that demonstrates optimal performance will be assessed for effectiveness using the mean absolute percentage error (MAPE). This metric will measure the divergence between the model's projected forecasts and the actual observed data for the corresponding timeframe. In the context of accepting or rejecting the null hypothesis, a model will be deemed "effective" if its MAPE stands below the 20% threshold for its 2022 test data forecast. Should an effective model be successfully developed, it will then be applied to produce a similar forecast for the 2023 home sales data.

Justification of Tools/Techniques:

Environment: Jupyter Notebook
Programming Language: Python
Python Libraries: Pandas, Matplotlib, SciKit-Learn, NumPy, StatsModels, Itertools, PMDarima and Prophet
Evaluation Metric: Mean Average Percentage Error (MAPE)

The evaluation of the optimized forecasting model's effectiveness will employ the mean average percentage error as the chosen metric. While mean squared error (MSE) and root mean squared error (RMSE) offer potential alternatives, their magnitude is subject to the characteristics of the study's units. Considering that nationwide home sales register in the thousands monthly, the MSE/RMSE values are likely to exhibit considerable absolute magnitudes. Within this context, the mean average percentage error (MAPE) emerges as a more standardized and intuitive measure of error. Operating within the conventional scale of 0 – 100%, MAPE delineates the percentage error between predicted and observed values. Assessing the mean average percentage error (MAPE) of a forecast does involve taking into account scenario-specific factors. However, a commonly recognized principle in the realm of business forecasting establishes that a "good" model is characterized by a MAPE below 20%, whereas a MAPE below 10% signifies a model of exceptional proficiency. Thus, when evaluating the null hypothesis, the efficacy of a model will be contingent upon whether its MAPE stays beneath the 20% threshold, signifying its status as a "good" forecasting model

Employing a comprehensive array of tools and techniques in Python, including Jupyter Notebook, Pandas, Scikit-Learn, Matplotlib, pmdarima, itertools, StatsModels, Numpy and Prophet, proves instrumental for crafting a potent time series forecasting model for home sales. Python's versatile ecosystem synergizes seamlessly with Jupyter's interactive interface, fostering iterative development and insightful documentation. Pandas' efficient data manipulation and transformation capabilities accommodate the intricacies of real estate data, while scikit-learn's diverse algorithms ensure robust modeling tailored to nuanced patterns. Matplotlib and Prophet facilitate visually intuitive data exploration and trend visualization, enhancing interpretability. The integration of pmdarima and itertools expedites preprocessing and feature engineering, while Statsmodels' refined statistical tools enrich model diagnostics. This arsenal of specialized tools and techniques collectively empowers data analysts to extract meaningful insights, construct accurate forecasts, and derive informed decisions within the dynamic realm of home sales.

Prophet, now called Prophet, is a Python library developed by Facebook for time series forecasting, particularly suited for home sales data. Its automatic trend detection, seasonality modeling, and holiday effects handling simplify modeling complexities. Prophet's user-friendly API allows rapid experimentation and fine-tuning of forecast parameters. Additionally, its built-in uncertainty estimation enhances result interpretation.
Scikit-learn proves its utility across a multitude of functions, encompassing tasks such as partitioning data into training and test groups, expediently computing mean squared error for forecasts, and various other essential tasks intrinsic to the analytical process at hand.
PMDarima brings benefits of automated ARIMA model selection and tuning. Its user-friendly API streamlines the modeling process, saving time and effort. pmdarima handles seasonality, trend detection, and hyperparameter optimization, leading to accurate predictions. This aids data analysts in focusing on interpretation and decision-making rather than manual model refinement. For home sales, pmdarima offers a convenient toolset to enhance forecast accuracy and aid strategic planning in a seamless, efficient manner.
Pandas' library is unique in its ability to provide adeptness in managing and manipulating sequential data. Its flexible data structures and functions enable efficient preprocessing and feature engineering, while seamless integration with other Python libraries simplifies analysis.

Project Outcomes: The project will generate a model that can forecast the volume of home sales through CY 2019 to 2022 with effectiveness demonstrated by a mean absolute percentage error (MAPE) of under 20% to reject the null hypothesis. If an effective forecasting model is generated, a forecast will also be performed to forecast the volume of home sales through CY 2023.

Projected Project End Date: 08/28/2023

Sources:

Housing data - Zillow Research. (2023, April 25). Zillow. https://www.zillow.com/research/data/

Olsen, S. (2023, February 11). *Zillow Home Value Index Methodology, 2023 revision: What's changed? - Zillow Research*. Zillow. https://www.zillow.com/research/methodology-neural-zhvi-32128/

Pierre, S. (2021). A guide to time series analysis in Python. *Built In*. https://builtin.com/data-science/time-series-python

Housing statistics and real estate market trends. (2012, January 13). www.nar.realtor. https://www.nar.realtor/research-and-statistics/housing-statistics-and-real-estate-market-trends

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 8/18/2023

Reviewed by:

Comments: Click here to enter text.