
PROJECT #2

OM 516 STEM/CREATE

Fall 2022

Topic: Forecasting

Problem Description

For a large majority of us, a home is the most expensive purchase of our lifetime. That being the case, we research at length not only the home we want to purchase but also the timing of the purchase. The whole real estate market is also big, roughly 16% of the U.S. GDP as of 2021. In this project, we focus on new listings, which directly impacts real estate operations. By using monthly data provided by Zillow.com between January 2018 and December 2021, we'll forecast via two alternative regression-based methods for the calendar year of 2022. For the first 8 of the 12 months, we will also calculate the accuracy of our forecast.

Methods

In this project you're asked to use two regression-based forecasting methods: 1) *Decomposition Method of Forecasting with Trend and Seasonal Components* and 2) *Multiple Linear Regression*. The methods are outlined in the following two subsections.

Decomposition Method of Forecasting with Trend and Seasonal Components

This forecasting method captures both seasonality and trend and generates forecasts that reflect the patterns observed in the past data. The accuracy of the method depends on the individual strengths of trend and seasonality components.

The method addresses seasonality and trend separately, and combines them in the final forecast. Steps of the method are:

1. Calculate seasonal indices
Steps 1-3 on Page 286 of the textbook outline the procedure to calculate a seasonal index for each month (i.e., season) of the year using data from the past four complete years (2018-2021).
2. Deseasonalize the past data
Divide raw data into seasonal indices. This takes seasonality out of the picture. And helps us see the trend more clearly.

3. Capture trend via simple linear regression
Compute intercept (a) and slope (b) by taking time as the only independent variable ($t = 1$ for January 2018, 48 for December 2021), i.e., compute prediction equation $Y = a + bt$.
4. Calculate preliminary forecasts by extending the trendline to the future
Apply the trendline (plug $t = 49, \dots, 60$ into the prediction equation) to the next 12 periods.
5. Calculate final forecasts by seasonalizing the preliminary forecasts
Multiply the preliminary forecast of each future month by its seasonal index.

A common confusion about the method is the need for Step 2. If Step 2 is skipped, trend can still be extracted in Step 3 but its quality will be affected.

Multiple Linear Regression

This forecasting method is purely causal. It captures both seasonality and trend via independent variables specifically for these purposes and generates forecasts from the result of multiple linear regression applied to the data. Steps of the method are:

1. Generate additional independent variables
 - a) Add time ($t = 1$ for January 2018, 48 for December 2021).
 - b) Add 11 dummy binary variables to separate the 12 months from each other. The 11 dummies can be named February?, ..., December?, basically one for each month other than January. Each dummy variable's value is determined by answering the question in its name, 'yes' is 1 and 'no' is 0. For January, all 11 dummies take the value 0. For each other month, only the dummy associated with that month is set to 1 and all others get 0.
2. Run multiple linear regression
Use new listings in your assigned area as the dependent variable and the 12 independent variables defined in Step 1 above build and run multiple linear regression. The outcome from regression is a linear function with an intercept and 12 slope coefficients (one for each independent variable).
3. Calculate forecasts
Populate the values of the independent variables for the 12 months of 2022, and then plug them in the prediction equation from Step 2 to calculate the forecasts.

Resources

- Data
 - All in a single ("Metro_new_listings_uc_sfrcondo_month.csv") file
 - Each team is asked to analyze one row of data from one metropolitan statistical area (MSA) as indicated in the second column ("SizeRank") as well as the row name in the Pandas data frame created therefrom. Note that team 0 is the instructor, and there are far more rows than teams.
- A Python library ("PublicFunctions.py") including functions you may need for the project.
- Python

- Install Anaconda (<https://docs.anaconda.com/anaconda/install/>)
- Install Jupyter Notebook (<https://jupyter.org/install>)
- An example demonstrating both methods in MS Excel
- Power Point presentation template

Tasks

1. Use the decomposition method to calculate the 2022 forecasts (30pts).
2. Use the multiple linear regression method to calculate the 2022 forecasts (30pts).
3. Compare the accuracy of the two sets of forecasts of the decomposition method thru mean absolute deviation (MAD) on the first 8 months of 2022 (8pts).
4. Discuss any improvements to the implementation of the decomposition method (8pts).
 - Are there any points in the method that didn't make good sense to you? Did you think of revising parts of the method? If so, how?
5. Discuss how the seasonality and trend are treated by the two methods (8pts).
 - Are they treated independently or is their interaction captured by the decomposition method?
 - Are they treated independently or is their interaction captured by the MLR method?
6. Discuss the fit of these two forecasting methods to the application at hand (8pts).
 - How did they perform in the 8 months of 2022? Are you happy with either method's performance overall?
 - Did Covid-19 have an effect on the performance of the two methods?
7. What other explanatory (independent) variables can you obtain and use to improve the forecasts (Extra credit up to 20pts)?
 - A couple of macroeconomic variables that pop into mind are labor force size and unemployment rate (both available at the MSA level from <https://data.bls.gov/PDQWeb/la>), mortgage rate (national average for 30yr. fixed mortgages are available from <https://www.freddiemac.com/pmms/pmms30>). Note that these are provided just as examples, many others can be thought of (and hopefully obtained at least at the national level but if you can find it at the MSA level that'd certainly be better).
 - To earn extra credit, you need to define the additional independent variables, obtain the data, incorporate them into either method (or both), and obtain improved forecasts.

Deliverables

- Python notebook(s).
- Power Point presentation (which you may present to class during our discussion).
- Project2 assignment on Blackboard where you can submit the two afore-mentioned files by the due date (11:59PM on 10/25/2022).
- Discussion in class (on 10/26/2022).