Ad Ease Website analytics Business Case

Website Views forecasting

Topic: Time Series

Duration: 1 Week

Why this case study?

From the company's perspective:

- Ad Ease is an ads and marketing-based company helping businesses elicit maximum clicks @ minimum cost.

- AdEase is an ad infrastructure to help businesses promote themselves easily, effectively, and economically

- Ad ease trying to understand the per page view report for different wikipedia pages for 550 days, and forecasting the number of views so that you can predict and optimize the ad placement for your clients.

- By leveraging data science and time series, Ad Ease can forecast page visits for different languages.

From the learner's perspective:

- Engaging with this case offers a practical understanding of how time series operate in ads and what are the challenges faced.

- Time series comprises of different algorithms and pre-processing techniques that can significantly improve forecasting.

- This exercise hones the participant's skills in dealing with missing values,modification in data, feature engineering, and model traning and forecasting.

- Additionally, learners gain hands-on experience in tackling real-world problems, transforming raw data into actionable insights that can guide business strategies.

DataSet Explanation

There are two csv files given

1. **train_1.csv** : In the csv file, each row corresponds to a particular article and each column corresponds to a particular date. The values are the number of visits on that date.

   The page name contains data in this format:

   SPECIFIC NAME _ LANGUAGE.wikipedia.org _ ACCESS TYPE _ ACCESS ORIGIN

   having information about page name, the main domain, device type used to access the page, and also the request origin(spider or browser agent)

2. **Exog_Campaign_eng**: This file contains data for the dates which had a campaign or significant event that could affect the views for that day. The data is just for pages in english.

There's **1 for dates with campaigns** and **0 for remaining dates**. It is to be treated as an exogenous variable for models when training and forecasting data for **pages in english**

What is Expected?

You are working in the Data Science team of Ad ease trying to understand the per page view report for different wikipedia pages for 550 days, and forecasting the number of views so that you can predict and optimize the ad placement for your clients. You are provided with the data of 145k wikipedia pages and daily view count for each of them. Your clients belong to different regions and need data on how their ads will perform on pages in different languages.

Submission Process:

Type your insights and recommendations in the text editor.

- Convert your jupyter notebook into PDF (Save as PDF using Chrome browser’s Print command), upload it in your Google Drive (set the permission to allow public access), and paste that link in the text editor.
- Optionally, you may add images/graphs in the text editor by taking screenshots or saving matplotlib graphs using plt.savefig(...).
- After submitting, you will not be allowed to edit your submission.

General Guidelines:

This scenario mirrors real-world challenges and embodies the tasks data scientists frequently grapple with. Embrace this opportunity to dive deep and simulate a professional experience.

During the course of this study, it's possible to face hurdles or even feel daunted:

- Re-evaluate the problem statement periodically to assure alignment with objectives.
- Deconstruct multifaceted tasks into simpler, achievable steps.
- If faced with code errors or issues, turn to online forums or official documentation. Problem-solving acumen is indispensable for data scientists.
- Collaborate with colleagues. Engaging in the discussion forum can offer diverse perspectives, aiding in overcoming obstacles or sparking new ideas.
- Revisit lectures or explore external resources for topics you're unsure about.
- For any overarching issues or if the problem statement appears ambiguous, don't hesitate to contact your Instructor.

Remember, every challenge faced is an opportunity to grow. Approach this case with enthusiasm, diligence, and an open mind.

What does “good” look like?

1. Define the Problem Statement and perform pre-processing and EDA

	Hint	Approach
1. Definition of problem	Start by understanding the problem statement. What's the objective of Ad Ease?	The main aim is to forecast multiple time series for different languages for their page visit.
2. Observations on Data	A thorough understanding of the dataset structure is key. Observe the shape of data, data types of all the attributes, etc.	Use functions like data.info(), data.describe(), and data.shape in Python.
3. Missing values	How to deal with NaNs ?	● Observe for NaN value in the data. ● Is there any pattern? ● Can we use mean to fill nans ? ● Can we use linear interpolation ? ● Can we drop missing values ?
4. Modification in the data	What kind of change do we want in our data for better analysis and forecasting?	Extraction of language code ● Using string manipulation ● Using Regex
5. EDA	Plot the distribution to get the insights	● Use matplotlib,seaborn etc to viz ● Explain the plots, what do you infer from them.
6. Aggregate and Pivoting	Why do we need to aggregate data? Why do we need to Pivot data?	Using group by and Aggregating ● Group by on what basis ? ○ On language basis because we need the forecasting per language ● Aggregating on what ? ○ We can aggregate on sum/ mean to get aggregate page visits value for a language on a particular date. ● Pivoting/transposing data would make dates as a single column ○ This makes our data in standard format
7. Time series plots	Observing time series plots for all languages	Once we have our data in standard format, we can plot data for all the

		languages and come up with observations.

2. Stationarity, decomposition, detrending, ACF, and PACF

	Hint	**Approach**
8. Stationarity test and decomposition	Why do we need stationary time series?	AdFuller test is the easy way to check Stationarity
9. De-trending and de-seasoning	What is the need to remove the trend and seasonality of our time series?	You can use differencing to detrend time series
10. Getting insights into time series characteristics	Using ACF and PACF plots	● Acf and Pacf will give seasonality and parameters for the models. ● Try PACF plot with original time series without de-trending and de-seasoning. ● ACF plot with de-trended & de-season time series

3. Model building and Evaluation

	Hint	**Approach**
11. Data splitting	Before model building split the dataset into training and test sets.	Try to keep the test set with multiple of seasonality.
12. ARIMA model	Forecast using ARIMA without exogenous variable	● Creating and training the Arima model with p and q from ACF and PACF plots

13. SARIMAX model	Helps with exogenous variable	● Use exogenous variable with sarimax
14. Facebook Prophet	Properly install this library. If not installing on local system, use colab.	● Create proper format of data for Facebook Prophet ● Use exogenous variable with Prophet
15. Comparison	Get the metric for all the languages	● The process for all other languages remain the same. ● Exogenous variable is not given for other languages ● Explore prophet method

4. Results Interpretation & Stakeholder Presentation

	Hint	**Approach**
a. Understand the Business Context	The primary concern for Ad Ease is to forecast the page visits	a. Understand Ad Ease objectives in forecasting page visits. b. Recognize the challenges faced by Adease and the factors affecting their performance.
b. Interpreting multiple time series forecasts	Understand multiple forecast With and without exogenous variable	a. Understand why page visit are more for English and not for others b. Understand the seasonality for every time series
c. Visual Representations	Visuals can often convey information more effectively than numbers alone.	a. Use plots to represent the distribution of page visits of different languages b. Showcase the ACF and PACF plots to get insight of time series

Questionnaire:

1. Defining the problem statements and where can this and modifications of this be used?
2. Write 3 inferences you made from the data visualizations
3. What does the decomposition of series do?
4. What level of differencing gave you a stationary series?
5. Difference between arima, sarima & sarimax.
6. Compare the number of views in different languages
7. What other methods other than grid search would be suitable to get the model for all languages?