Model Setup: Prophet Univariate Time Series Forecasting Algorithm

We utilize the prophet model to model the historical trend of university application status in the United States and evaluate the model prediction accuracy. The prophet model, which Meta develops, is the forecast model that can fit historical time series data and predict the future. We chose this model because it can handle the seasonality in the dataset.

In our dataset, users posted their application status with the decision date and in the following types: interview, accepted, rejected, waitlist, and others. To analyze the historical trend, we count the number of status updates per date from 2006 to 2020, fit the model with each type of decision update and make predictions. Notes that we did not analyze the classes: "waitlist" and "other" because of a lack of data points.

Moreover, we observed many data vacancies in some months, which was caused by the seasonality of our dataset. The vacancies lead to inaccurate predictions of these months. To improve the prediction accuracy, we perform interpolated the test dataset to assign a value to every date that doesn't have data. In addition, we compared various interpolation methods by mean square error (MSE) and visualized the prediction result.

We observed that the number of status changes is extremely volatile in the fixed season. By contrast, in other seasons, the number is small and steady. And the prophet model cannot handle the dramatic difference between these two types of seasons. Hence, we took five days' mean to the test dataset and evaluated the prediction result by MSE and visualized chart.

Results & Analysis

1. Predictions of different type of application status.

In this section, we show the prediction results of the number of status changes for the following types of application status: all, interview, accepted, and rejected.

Table x shows the MSE of each type of application status. Figures x, x, x, x compare the ground truth and model predictions of different types of application status, including all, interview, accepted, and rejected.

a. For the slack season, the model makes a prediction that is obviously wrong because the dataset lacks data points in the slack season.

b. For the peak season, the predicted value is roughly the mean of actual values. In other words, the model cannot fit the extremely high-frequency change of the number of status changes in peak season.

To fix the first issue, we perform interpolation to our dataset so that all dates in the time series have value either from the dataset or interpolated. We expected better prediction accuracy in the slack season with interpolation. To fix the second issue, we calculate five days mean of the original dataset and use the five days mean to fit the model. We expect that we can remove the noise from the dataset by using five days mean instead of the original values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Application status | All | Interview | Accepted | Rejected |
| Mean square error | 654.02 | 141.54 | 46.63 | 99.32 |

一張含有 圖表, 文字 的圖片

自動產生的描述

一張含有 文字, 圖表, 螢幕擷取畫面 的圖片

自動產生的描述

一張含有 圖表, 螢幕擷取畫面 的圖片

自動產生的描述

一張含有 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述

1. Prediction of different interpolation methods

In this part, we show the prediction results of the number of status changes by performing the following interpolation methods to the test dataset: linear, backfill, and quadratic, and compare it to the prediction without interpolation.

Table x compares the MSE of different interpolation methods. Figure x shows the original prediction result, and figures x, x, and x offer the result by linear, backfill, and quadratic interpolation, respectively.

After performing each interpolation method, the mean square error is much higher than the original result. And all interpolation methods give nearly the same prediction result. However, by comparing the figure of the original prediction and the prediction after interpolation, we observed that the prediction for slack seasons become much more accurate after interpolation. Therefore, we can conclude that decreasing the prediction accuracy of peak season is a tradeoff for better prediction accuracy during the slack season for the model. And we believed that it is necessary to perform interpolation for the dataset to avoid prediction errors in the slack seasons, and we interpolated data with the backfill method in the next section because it is the most straightforward approach.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Interpolation method | No interpolation | linear | backfill | quadratic |
| Mean square error | 654.02 | 777.46 | 777.83 | 772.53 |

一張含有 圖表, 文字 的圖片

自動產生的描述

一張含有 文字, 圖表 的圖片

自動產生的描述

一張含有 文字, 圖表 的圖片

自動產生的描述

一張含有 文字, 圖表 的圖片

自動產生的描述

1. Prediction of taking five days mean.

Recall that we want to reduce noise in the peak season. Hence, in this section, we make a prediction of the number of status changes per day that taking a five-day mean and compare it with the prediction that do not taking a five-day mean. Notes that we apply backfill interpolation to both datasets before fitting the model. Table x compares the MSE of prediction with and without a five-day mean. Figure x shows the prediction with a five-day mean, and Figure x shows the prediction without a five-day mean.

The table shows that the MSE increase if taking a five-day mean. However, by comparing the prediction curve, we observed that the prediction curve is smoother and has less noise, especially in the slack season.

|  |  |  |
| --- | --- | --- |
|  | Five-day mean | No five-day mean |
| Mean square error | 874.05 | 777.83 |

一張含有 文字, 圖表, 螢幕擷取畫面 的圖片

自動產生的描述一張含有 文字, 圖表 的圖片

自動產生的描述

Conclusion

The model can fit the seasonal change and make a prediction. However, the model cannot handle the rapid value change in peak season, and a significant error occurs while modeling the slack seasons.

We can improve accuracy in the slack seasons by interpolating the missing value in the dataset, but the overall mean square error also increases.

We can decrease the noise of prediction value by taking the five-day mean to the dataset, but the overall mean square error also increases.

In the future, we can explore better metrics to evaluate the prediction accuracy to replace the mean square error metric.