Course: Programming for Business Analytics
Assignment: Airline Demand Forecasting
Submitted By: Tarun Shrivas, 2nd Member

This Python program provides robust demand forecasts for airline bookings, given the flight departure date (*departure_date*) and the booking dates (*booking_date*). The program aims to optimize airline bookings forecast and minimize forecasting errors. The program uses historical flight bookings data (*trainingDataFile*) and estimates the final demand (*final_demand*) using the best of four mathematical models used in the program. The program uses naïve forecasting(*naïve_forecast*) data (*validationDataFile*), as a benchmark to select the best forecasting model. For this, the MASE (mean average scaled errors) of each forecasting model is compared with the MASE of the naïve forecasts. The program selects the best forecast model and uses that to forecast the final demand, for all the days up to the flight's departure date.

We used *"airline_booking_trainingData.csv"* .csv file as input for *trainingDataFile* and the *"airline_booking_validationData_revised.csv"* .csv file for the *validationDataFile*. We read these .csv files to create our training data frame *train_dat* and our validation data set *valid_dat*. For developing our forecast models, we have mutated the existing variables to create new variables. Foer example, we mutated *days prior = departure date – booking date* to create *days_prior* columns in *train_dat* as well as *valid_dat*. We also *created a week_day column for which day of the week for the departure_date was taken.* Likewise, created other variables required to be used in one or more of our forecast models. Table 1 given below provides a brief snapshot of the four forecasting models that we developed in our program. All the models used the same datasets for training and the same dataset for validation. All the models predicted *final_demand* for a given *departure_date* and *booking_date*. The program worked out the MASE error for each of the four models.

Table 1. Forecasting models used in the program

MODELS	LOGIC USED	VARIABLES USED	FORECASTING FORMULA =
Model_A	Multiplicative	***days_prior, week_day***	*= cum_booking/(average_booking_Rate)*
Model_B	Multiplicative	***days_prior***	*= cum_booking/(average_booking_Rate)*
Model_C	Additive	***days_prior, week_day***	*= cum_booking + average_remaining_demand*
Model_D	OLS Regression	***days_prior, week_day, cum_bookings***	*= predict(final_demand)*

Model_A is a multiplicative model and uses *days_prior* and *week_day* as inputs to work out an *average_booking_rate* for data grouped by unique combinations of *days_prior* and *week_day.* Dividing the *cum_bookings* by *average_booking_rate* provided us the demand forecast. Model_B was developed with the similar logic however, it used only the *days_prior* variable as an input. Model_C is an additive model and adds the *average_remaining_demand* to the *cum_bookings*, for the *final_demand* forecast. Model_D is using an OLS regression model to estimate the *final_demand* as dependent variable and days*_prior*, *week_day*, *cum_bookings* as independent variables. The predicted *final_demand* values from the model are used as forecast.

For all these models, we calculated the Mean Absolute Scaled Error (MASE). Our program selects the model having the lowest MASE value. Once a model is selected, our program returns the summary information on all the model, and a forecast data frame having *departure_date*, *booking_date,* and *robust_forecast* as its variables. The program also writes the output forecast data frame to a .csv file viz. *"airline_robust_forecastData.csv"* file.