# Public Project Summary: Bruno Helmeczy

This Technical Discussion Report describes my completed Capstone Project, the final graduation requirement for earning the title of Master of Science in Business Analytics, from the Business & Economics Department of the Central European University. During this Capstone Project, students work for external clients as consultants, looking to create added value for client organizations via an end-to-end analytics project, utilizing skills, concepts & tools learned throughout the curriculum. My technical discussion report is divided into 4 Chapters: Context & Analytics Plan, Downloadable Standardized Report, Improving Rooms Sales Forecasting with S-ARIMAX models, & Added Value & Recommendations.

In Chapter 1 I introduce my client organization, Archipelago International, explore their problem statement, and formulate an analytics plan to improve their performance through analytics. My capstone project revolves around improving Archipelagos' analytics-platform, a Shiny-based web application delivering data-based insights to commercial leaders at both the headquarter-, and property-level. I found two opportunities to add value to the analytics platform:

1st Designing a parameterized Report Template, summarizing key information hotel leaders may need, & illustrating how to make it available for a selected hotel from the Shiny-based analytics platform. 2nd implementing time series-based forecasting models, testing their viability versus current models, hoping to improve Archipelago Internationals' forecast accuracy. Chapters 2 &3 discuss the technical details surrounding the creation of said parameterized report

# Parameterized Report:

Throughout designing the report template, I followed a business-oriented approach, from a hoteliers' perspective. I looked to answer common commercial performance-related questions, which arise periodically, most frequently during weekly Revenue Strategy meetings & discussions. The charts & summary tables I coded look to answer these questions, while providing visual cues, where further investigation may be needed, ideally via the analytics platform.

The high-level questions my charts & tables looked to answer: 1st How are hotels performing versus Last Year?; 2nd What changed since we met Last Week?; 3rd How does the next 3 month look like versus Last Year?; 4th How many rooms can we expect to sell & how far in advance of actual arrivals?; & 5th How confident is the data science / analytics team in their forecasts for the hotel in question?

Once I wrote the R codes for the plots & tables used to answer these questions, I prepared a static R Markdown Report. The next step then was to dynamically change the data used for generating these plots & tables, which I achieved by means of introducing meta-parameters in the Markdown Report Template YAML header, using this newly-created parameter as filter arguments to subset my relevant data, given the parameters' value.

Being able to dynamically change the R markdown files final outputs, my next step was to render the markdown file remotely, along the way determining the final output files' format, as well as the parameter value that should be used while rendering the markdown file. Finally, I built a Minimum Viable Shiny Application through which I can illustrate how a report download may be executed. To do so, I created a most-simple user interface, solely with a dropdown menu of available hotel names to select from & a download button initiating the report to render, and subsequently download the R markdown output file, with the default filename being of the hotel-name-plus-todays-date format.

## Forecast Modelling

My motivation to create room sales forecasting models for Archipelago International are despite the fact that they employ well-performing Machine Learning Models. Conceptually however, not employing traditional, or even more advance, time series methods seemed unsatisfying, & I wanted to find out, why such models were left out thus far. Furthermore, this project deliverable allowed me to challenge myself not only with implementing my chosen models with solid statistical foundations, but do so in a reproducible manner for ca. 150 properties.

My goal was to fit, cross-validate, implement, & test multiple models, which rival the performance of currently employed models of Archipelago International. I compared all models performance by assuming forecasts are to be generated on every day of December 2020, forecasting 91 days ahead on each day. These forecasts were my test set outputs, allowing forecast error statistics comparisons.

To be able to characterize models' performance across the entire hotel-chain however, I could not simply use common error statistics, like Mean Absolute Error across all hotels, as some had less then 15 rooms, others over 400. Hence, I chose to characterize models performance by rankings based on mean absolute forecast error within a given hotel, doing so both via calculating mean absolute error from all forecasts made for all forecast horizons, as well as grouped by weeks before arrival.

This supports the business notion, that how important a forecasts' inaccuracy is, quite strongly depends on how many days, weeks, months ahead is the given forecast is being prepared. Once aggregating all model rankings across all hotels, I could derive frequency-based statistics, answering questions such as: 'For what percentage of hotels does model XYZ rank 1$^{st}$, considering the complete forecast horizon?'; or 'For what percentage of hotels does model ABC rank 1$^{st}$, when forecasting within one week of arrival?'

Throughout this model-building-slash-forecasting exercise, I built 3 models of gradually increasing complexity: 1$^{st}$ a Seasonal Autoregressive Integrated Moving Average Model; 2$^{nd}$ a Seasonal Autoregressive Integrated Moving Average Model, including indicator explanatory variables for holidays, and weekends; and 3$^{rd}$ a Seasonal Autoregressive Integrated Moving Average Model, also including room reservation-related variables, i.e. Reservations-on-Hand at multiple chosen number of Days Before Arrival, & Rooms Reserved (Picked Up) between these chosen Days Before Arrivals. The 2$^{nd}$ & 3$^{rd}$ model are also called Dynamic Regression Models, first applying a multiple linear regression on the target variable time series with the explanatory variables, then modelling these regressions' residuals as Seasonal Autoregressive Integrated Moving Average Series.

To deploy these models efficiently, I wrote 8-12 functions in R, helping me to both conceptually & functionally separate stages of the complete forecasting procedure, applying it 1 hotel at a time: 1$^{st}$ I loaded the data from relevant data tables, 2$^{nd}$ I imputed missing values, 3$^{rd}$ I created fold indices for cross-validation, 4$^{th}$ I fitted & selected models based on cross-validated forecast errors, 5$^{th}$ I forecasted the test data set; & 6$^{th}$ I aggregated forecast errors to enable comparing my models with Archipelagos' currently existing ones. Resultingly, in the forecast approaches final implementation, I simply 'pipe' these functions along & pass a hotels' name to execute the complete forecast procedure.

As the procedure was to be applied to not-so-small chain of hotels, I also had to deal with running my models within the above procedure efficiently. Hence, I also implemented parallel processing, though not with complete success. Eventually, my final model yielded the best results across the hotel chain, ranking 1$^{st}$ for 33 of Archipelagos hotels.

## Results & Added Value Estimation

Having ran my models & summarized their performance, I wanted to estimate the potential added value of my proposed models, based on Lee's (1990) premise: A 10% improvement in forecast accuracy translates to +0.-5%-3% revenues, but only during High Season. To do so, I made a number of assumptions, noted below. In essence, I assume the hotels where either of my models ranked 1st would have used the best of my models, whereas before they were using the best of their old models.

1. Before my models were proposed, Archipelago used the best model at their disposal, the model that ranked 2nd throughout my test set.
2. Error measures from my test set represent the true model errors, & as such the measured improvement in forecast accuracy represents the true improvement in forecast accuracy.
3. Archipelago used 1 of my models for the hotels for which either model ranked 1st considering the complete forecast horizon based on the test set data.
4. Current Euro to Indonesian Rupiah Foreign Exchange Rate (17332 to 1, as of 6th June, 2021) to have been the same since 2019 & will also hold in the future.
5. That High-Season in Indonesia covers the months starting May & ending in September.

Then, to derive a revenue range, I simply calculated each hotels' High-Season Revenues in 2020, as well as the improvement in forecast accuracy with my models versus their best model, and applied the 0.05% & 0.3% growth factor per 1% forecast accuracy improvement. Thus the product of the growth factor, the percentage improvement in forecast accuracy, and the hotels High-Season room revenues represent the upper- & lower bound of interval in which my models added value resides, with the assumptions above holding true.

Due to 2020 being an exceptionally weak year for hotels, I repeated the above exercise using 2019 data, yielding an interval of potential added value in both a good- & a bad year. Though the resulting added value figures (calculated in Euros), are not of material difference to the hotel-chain, they represent meaningful sums for individuals.

Finally, also based on these results, I provide a number of recommendations to improve the presented models further. Some of these include constraining forecasts with upper- & lower bounds, as well as utilizing 'Partial-Bookings data', i.e. all available data for a hotel at a given point in time, not just of 'Completed Stay Nights', i.e. data only from past stay dates.