Matthew Voynovich

Data Analytics

Professor Eleish

14 December 2025

Modeling Tourism’s Effect on Europe’s Economy

Abstract

Tourism is a major driver of economic activity across Europe, contributing significantly to employment, investment, and especially income. Understanding how tourism factors relate to economic growth is therefore essential to make informed decisions as policymakers and economists. This project examines the relationship between tourism-related variables and GDP across European countries using machine learning models. The initial hypothesis investigated was that an increase in tourist arrivals and related tourist expenditures are associated with an increase in the country's GDP. To investigate this, models were made to predict the GDP growth rate based on tourism statistics and control variables like population and exports. Results of these models were verified with R^2 scores and with feature importances to make sure tourism factors play a big role in predicting GDP growth. To further this investigation, models were then trained on a per country basis and their feature importances were investigated to discover which countries were the most sensitive to tourism in regards to their GDP growth. Using results from both of these models, conclusions can be made on the impact of tourism on Europe's GDP as a whole and the impact of it on a per country basis.

Data Description and preliminary analysis

The Dataset for this project was sourced from World Bank. World Bank is an international financial institution who collects and publishes global data such as GDP, poverty rates, tourism indicators, and health stats for countries in order to offer policy advice to help governments design effective economic policies. This project uses country GDP, tourism receipts, tourism arrivals, population, unemployment rate, and more in order to predict with a combination of tourism factors and control variables. Example rows inside the dataset can be found in Figure 1 below. On top of the features taken from World Bank, GDP growth and tourism growth were both features constructed using the data across rows to provide a growth percentage for both gdp and tourism from the last year. Documentation for this data can be found at https://data.worldbank.org/. The indicator codes used to retrieve this data are 'NY.GDP.MKTP.KD', 'ST.INT.RCPT.CD', 'ST.INT.ARVL', 'SP.POP.TOTL', 'SL.UEM.TOTL.ZS', 'NE.GDI.TOTL.ZS', 'NE.TRD.GNFS.ZS', 'FP.CPI.TOTL.ZG', 'PA.NUS.FCRF', 'IS.AIR.PSGR' which are all associated with the columns in figure 1 (besides our created GDP growth and Tourism growth features).

	country	year	GDP_constant_USD	Tourism_receipts_USD	Tourism_arrivals	Population	Unemployment_rate_pct	Gross_fixed_capital_formation_pctGDP	Trade_openness_pctGDP
15	Albania	2008	9.861658e+09	1.850000e+09	1420000.0	2947314	13.060	33.305671	75.248547
14	Albania	2009	1.012701e+10	2.013000e+09	1856000.0	2927519	13.674	32.440459	73.321358
13	Albania	2010	1.042810e+10	1.778000e+09	2417000.0	2913021	14.086	32.492623	75.532533
12	Albania	2011	1.068500e+10	1.833000e+09	2932000.0	2905195	13.481	33.837609	80.698999
11	Albania	2012	1.079016e+10	1.623000e+09	3514000.0	2900401	13.376	29.715028	76.968358
...	...	...	...	...	...	...	...	...	...
104	Switzerland	2015	6.941182e+11	2.014000e+10	9305000.0	8282396	4.801	25.074863	117.138482
103	Switzerland	2016	7.084773e+11	1.978700e+10	10402000.0	8373338	4.918	25.561928	123.565925
102	Switzerland	2017	7.181325e+11	2.039900e+10	11133000.0	8451840	4.797	26.044314	123.621807
101	Switzerland	2018	7.386743e+11	2.129400e+10	11715000.0	8514329	4.713	25.734453	124.428504
100	Switzerland	2019	7.471098e+11	2.125700e+10	11818000.0	8575280	4.394	26.437793	124.127612

	Inflation_pct	Exchange_rate_local_to_USD	Air_transport_passengers	GDP_growth	Tourism_growth
15	3.320871	83.894604	2.436910e+05	0.069071	0.250845
14	2.266922	94.978120	2.312630e+05	0.026908	0.088108
13	3.626047	103.936667	7.685330e+05	0.029732	-0.116741
12	3.429123	100.895833	8.297789e+05	0.024635	0.030934
11	2.031593	108.184167	8.143397e+05	0.009841	-0.114566
...	...	...	...	...	...
104	-1.143909	0.962381	2.701176e+07	0.016446	-0.057028
103	-0.434619	0.985394	2.585992e+07	0.020687	-0.017527
102	0.533788	0.984692	2.673257e+07	0.013628	0.030929
101	0.936335	0.977892	2.885799e+07	0.028604	0.043875
100	0.362886	0.993709	3.033965e+07	0.011420	-0.001738

Figure 1: Example entries in dataset

Data Description and preliminary analysis

With the gathering of our data, we move onto exploratory data analysis. Here we write an R script to generate statistics for our numerical columns and generate distribution plots for all of our columns.

```
    country         year      GDP_constant_USD    Tourism_receipts_USD Tourism_arrivals
Albania : 24   Min.   :2000   Min.   :6.153e+09   Min.   :4.900e+07    Min.   :    99000
Austria : 24   1st Qu.:2006   1st Qu.:3.880e+10   1st Qu.:2.070e+09    1st Qu.:  3376500
Belgium : 24   Median :2012   Median :2.018e+11   Median :5.632e+09    Median :  8001000
Bulgaria: 24   Mean   :2012   Mean   :5.198e+11   Mean   :1.065e+10    Mean   : 23481970
Croatia : 24   3rd Qu.:2017   3rd Qu.:4.673e+11   3rd Qu.:1.190e+10    3rd Qu.: 28070500
Cyprus  : 24   Max.   :2023   Max.   :3.702e+12   Max.   :7.252e+10    Max.   :217877000
(Other) :648                                       NA's   :241          NA's   :145
  Population       Unemployment_rate_pct Gross_fixed_capital_formation_pctGDP
Min.   :  281205   Min.   : 1.874        Min.   : 9.064
1st Qu.: 3006752   1st Qu.: 5.037        1st Qu.:20.540
Median : 7403584   Median : 7.326        Median :22.915
Mean   :16029408   Mean   : 8.964        Mean   :23.524
3rd Qu.:11388132   3rd Qu.:11.009        3rd Qu.:25.951
Max.   :83901923   Max.   :37.320        Max.   :53.714

Trade_openness_pctGDP Inflation_pct    Exchange_rate_local_to_USD Air_transport_passengers
Min.   : 21.11        Min.   :-4.448   Min.   :  0.3117          Min.   :      857
1st Qu.: 72.20        1st Qu.: 1.117   1st Qu.:  0.8039          1st Qu.:  1185479
Median : 95.93        Median : 2.207   Median :  1.0502          Median :  4909100
Mean   :105.92        Mean   : 2.990   Mean   : 23.5895          Mean   : 19082328
3rd Qu.:134.80        3rd Qu.: 3.581   3rd Qu.:  8.0839          3rd Qu.: 22931166
Max.   :256.40        Max.   :34.477   Max.   :372.5958          Max.   :170161848
                      NA's   :33                                 NA's   :123
```

Figure 2: Statistics for columns

Looking at the generated statistics, the scale of our features are all wildly different with some having a max value in the trillions and others having a minimum value in the tenth decimal place. This goes to show that if we used models that were sensitive to the scale of features, we would have to preprocess the data and scale them accordingly to get accurate measurements and predictions. However, since we are mainly using tree models we do not have to worry about scale of the features and therefore it does not require any preprocessing in that regard. We still have to preprocess our data by removing NA's as some of our models like Random Forest give warnings/errors. Performing this metric we go from 792 entries to 411 entries (as we also do not have the growth features we created if the previous year did not have a gdp or tourism receipts

entry). This happens because a lot of countries do not have their reported tourism statistics as we can see with a large number of NAs for the tourism receipts and tourism arrivals columns.

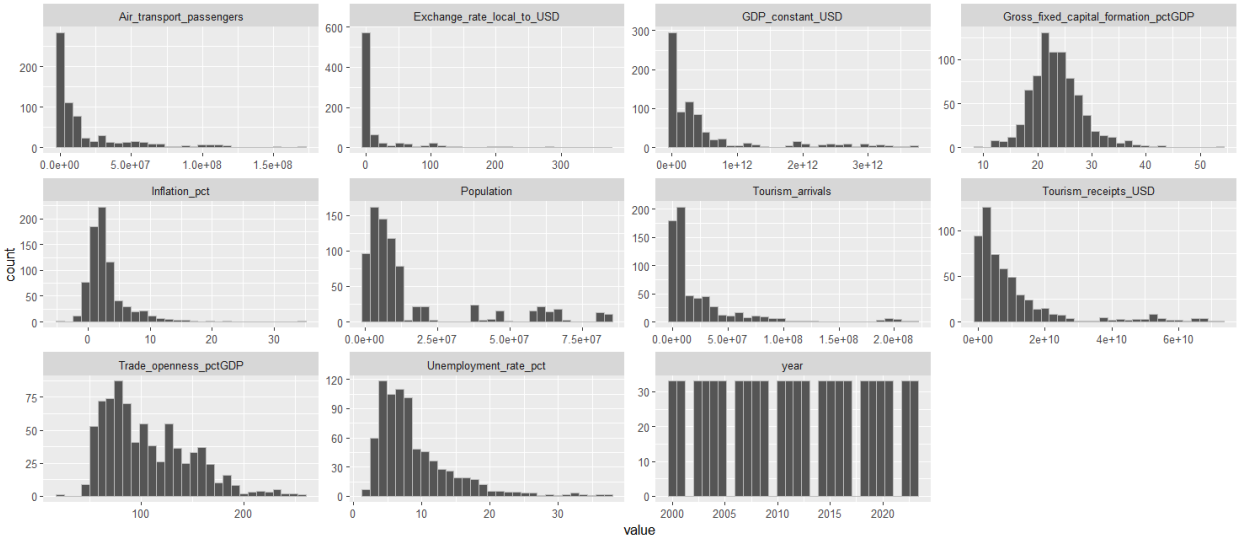

Figure 3: Distributions of numeric columns

Looking at the distribution of our variables in figure 3, we can see that the numeric values are typically right skewed. This makes sense as for each variable there are typically a few countries that have particularly high statistics such as France and Germany for air transport passengers due to them having large airports, or Southern Europe for tourism, and big countries like Germany or the UK for GDP. The more normally distributed features like Gross fixed capital formation percent GDP and trade openness percent GDP also make sense as they are typically percentages and Europe has an equal spread across these values. These skews can influence the models performance but due to trees being able to fit non linearly, the models should be resilient to the scale and skews without much preprocessing.

From here we then move on to check the correlation of the tourism features with GDP. We plot them against each other alongside doing a correlation calculation.

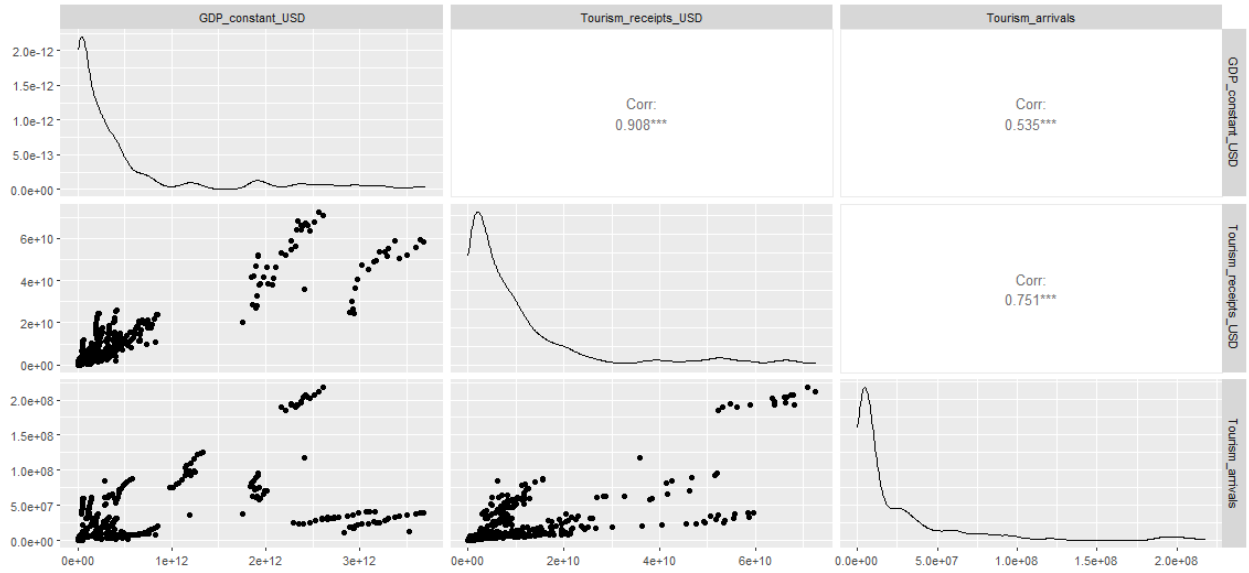

Figure 4: Tourism features correlation with GDP

As we can see, tourism is strongly correlated with GDP. Despite their being slight noise with their relationship, there is a positive correlation that approximately fits on a best fit line. This suggests that our models should be fairly strong in predicting GDP growth using tourism features alongside the control variables.

After doing EDA we can see the distributions of our features and how they correlate with our GDP target. This opens our eyes to potential sources of data like the dropping of rows from 792 to 411. Since our dataset was already fairly small, we were very likely to overfit the data to start with and are even more likely now. Additionally, though we are using trees so the distributions and scales do not matter so much, there may be issues with preprocessing the data that cause our model to perform worse than otherwise.

Model Development and Application of models

To start answering our hypothesis, we first start by solving our first investigation. We train two separate models off of the full Europe dataset using Tourism_receipts_USD', 'Tourism_arrivals', 'Population', 'Unemployment_rate_pct', 'Gross_fixed_capital_formation_pctGDP', 'Trade_openness_pctGDP', 'Inflation_pct', 'Exchange_rate_local_to_USD', 'Air_transport_passengers', and 'Tourism_growth' as features and our created feature ‘GDP_growth’ as the target. The two models chosen are Random Forest and Light GBM. Both models are ensemble models that combine multiple decision trees to make a final decision. These models were chosen as they do not require scaling for their data and they reduce overfitting through the type of ensemble. Random Forest is a bagging ensemble that trains multiple trees on random samples in parallel to then combine their predictions into one final vote. Light GBM and gradient boosting models in general are boosting ensembles that sequentially train models by increasing the weights of points that the previous model got wrong to correct mistakes. The model then takes the vote from all the models to make a final prediction. These two models were evaluated with a train test split of 80-20 using R^2 score and Root Mean Squared Error (RMSE). Both the random forest models and the Light GBM models were trained with 500 trees and stock settings/hyperparameters otherwise.

	Random Forest	LightGBM
R^2 Score	0.443224	0.393787
RMSE	0.030866	0.032207

Figure 5: Full Europe Data Model Results

As we can see the best performing model ends up being our random forest model. This is likely due to bagging being better at preventing overfitting than boosting. Both models do perform fairly well with a R^2 of 0.44 for Random Forest and 0.39 for Light GBM. While this R^2 might be low for predicting the GDP number itself, this is a good number for GDP growth as GDP growth is noisy and volatile because it is very dependent on other macroeconomic factors that we may not have as part of our control variables.

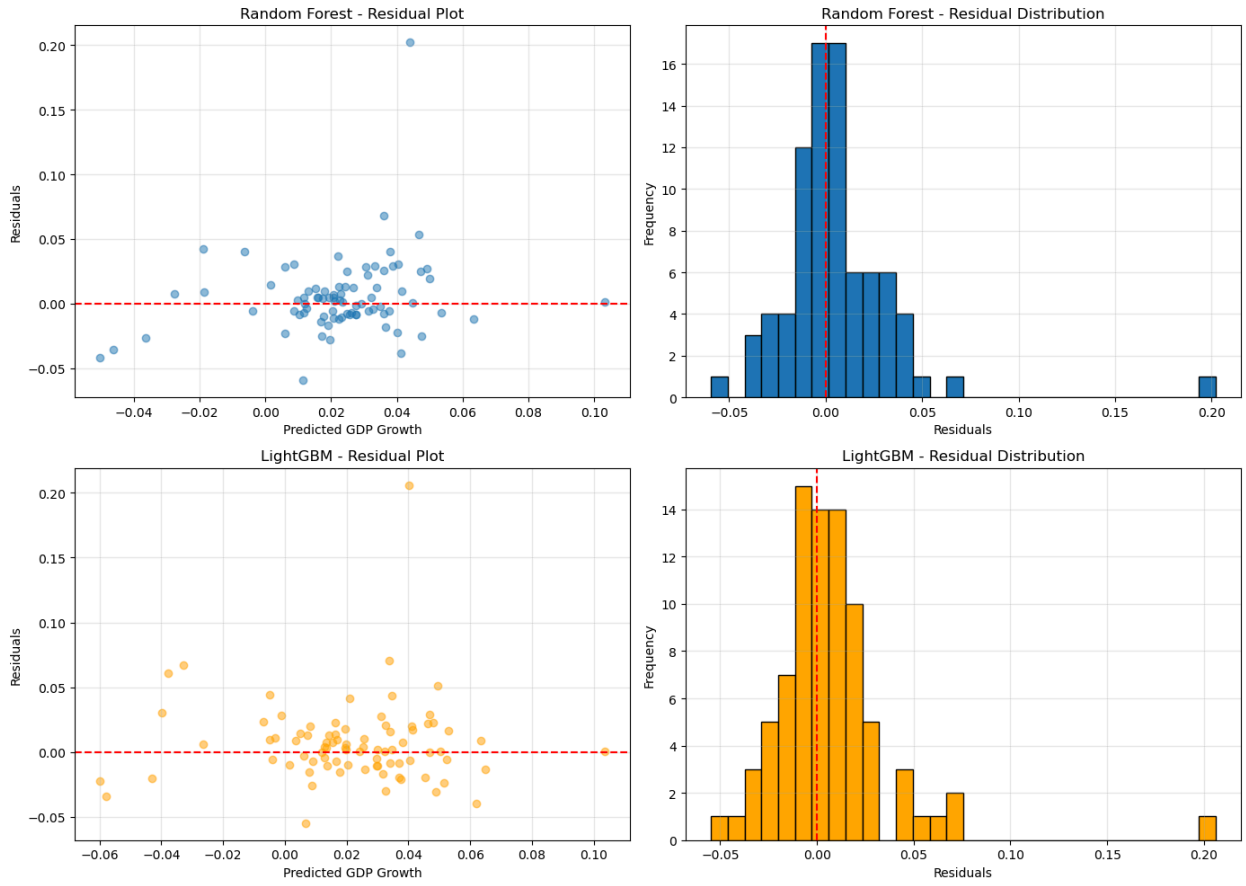

Figure 6: Residual plots for Random forest and Light GBM models predicting overall Europe GDP growth

The RMSE of the models is also very good indicating that the residuals typically predict within 3% of the GDP Growth. We can confirm this by looking at the residual plots in figure 6. The models show a tight spread on the predicted values and the residuals where they are

clustered around the x-axis meaning small residuals. Then by looking at the distribution of the residuals, we can see that they are centered around 0 and have a tight normal distribution with a singular point that is an extreme outlier. This is likely what drives the root mean squared error so high. Thus the mean absolute error would likely be lower than the 3% figure that we get from RMSE. These plots and the model performance statistics go to show that the models are fairly good at predicting GDP Growth off of tourism factors and control variables.

Looking at figure 7, we can see that the top feature for both models is tourism growth. For the Random Forest model specifically, tourism growth accounts for 40% of the decision making in the model. Light GBM does not do percentages but the feature is also the top for it. The features for light GBM are much closer in relevance but this is likely due to the nature of boosting where it tries to correct errors and therefore other features that may not be as important in a single decision tree become more important overall. An interesting thing is that tourism arrivals are a bad performing feature while tourism receipts are around middle of the pack, with tourism growth accounting for the majority of the prediction importance in regards to tourism indicators. While this is surprising given the importance of tourism growth as a feature, it also makes sense as tourism growth is a percent and therefore correlates better with gdp growth. Because tourism receipts and arrivals are just numbers it is harder to get their true importance in predicting GDP growth as the predictions are made without past GDP level and past tourism values to put them in context. These feature graphs confirm that tourism is a relevant factor in predicting the GDP growth of europe countries and with the models performing fairly well for a noisy target in GDP growth, we can measure GDP growth fairly accurately based on tourism indicators. With positive correlation between GDP and the tourism indicators and our models performing well on predicting Europe’s overall GDP, this indicates that an increase in tourism does indicate and predict an increase in the country’s GDP.

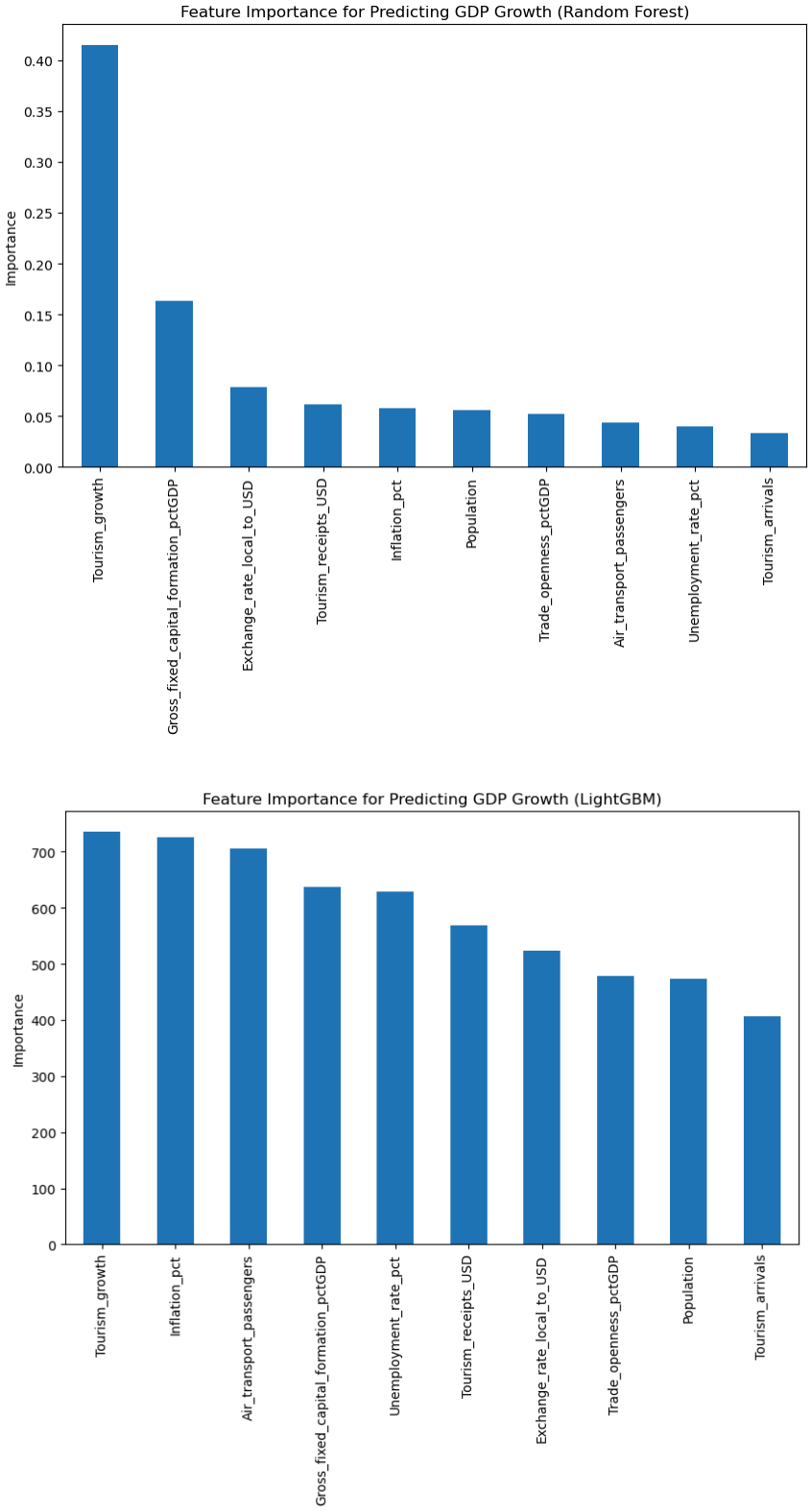

Figure 7: Feature Importances for Random Forest and Light GBM predicting overall Europe GDP Growth

Next, we move on to our second investigation to investigate the impact of tourism on the country level for GDP Growth. We do this by training Random Forest and XG Boost models for each country. XG Boost is another type of boosting algorithm that we use since it has definitive percentages for feature importances compared to Light GBM. They are then validated using time series cross validation that trains on past data and predicts on the future. Time series cross validation is done to see the model's true accuracy and to combat overfitting due to the small amount of data. Additionally, to prevent severe overfitting due to each country having about 20-30 entries, we train each model with only 50 trees and a small learning rate for XG Boost. These models were evaluated with an R^2 score to see how well they were actually performing. Some models were missing accurate R^2 scores as after preprocessing they did not have enough entries to do an accurate time series cross validation.

	RF_Tourism_importance	XGB_Tourism_importance	RF_R2_CV	XGB_R2_CV
Poland	0.502242	0.558069	-3.985423	-1.769850
Slovenia	0.443339	0.308877	-3.487811	-2.037817
Romania	0.442711	0.100876	-1.692953	-3.322596
Czechia	0.310263	0.135080	-6.545667	-31.857294
France	0.289056	0.034749	-20.956845	-6.785779
Portugal	0.287374	0.144327	-4.397474	-3.286806
Ireland	0.273203	0.260884	-3.037209	-10.138649
Finland	0.228167	0.474074	-16.979362	-21.613793
Cyprus	0.218478	0.095447	-34.435491	-12.409146
Croatia	0.215423	0.047491	-17.859059	-20.552853
Albania	0.215224	0.478468	-16.191330	-21.186263
Switzerland	0.192102	0.131877	-9.947053	-15.349026
Germany	0.153579	0.023994	-25.783456	-30.327508
Italy	0.099956	0.179766	-5.801774	-16.055896
Belgium	0.082131	0.050989	-3.418195	-5.299789
Serbia	0.070367	0.058779	-13.712131	-32.174818
Hungary	0.065430	0.001252	-1.103490	-0.611763
Bulgaria	0.042398	0.061693	-26.630017	-20.739421
Austria	0.038974	0.034641	NaN	NaN
Estonia	0.033298	0.018800	NaN	NaN
Greece	0.032636	0.022584	-40.268172	-49.545902
Netherlands	0.032134	0.220147	-19.016743	-8.183785

Figure 8: Per Country Model Results

As shown in figure 8, the time series cross validation R^2 scores for all countries are extremely bad indicating a lack of predictive power for tourism on a per country basis. This shows that tourism indicators have a contemporaneous relationship with GDP. This reveals that the models should not be used to predict GDP growth for the future, but it does not discount the correlation of tourism in relation to a country's GDP growth. Thus, the feature importances are still able to tell us how strongly tourism growth correlates with GDP growth and what countries are most sensitive to tourism regardless of model accuracy.

Figure 9 reveals which countries are most sensitive to tourism and therefore which ones have the greatest correlation between GDP growth and tourism. According to both models, Poland is the most sensitive country to tourism in regards to their GDP growth. From there the models have some differences but Slovenia, Ireland, Romania, and Finland are all within the top ten for both models. These models despite being bad predictors of GDP growth still serve an important purpose in telling us which countries GDP Growth strongly aligns with their tourism growth and therefore serves to influence further investigations into these relationships.

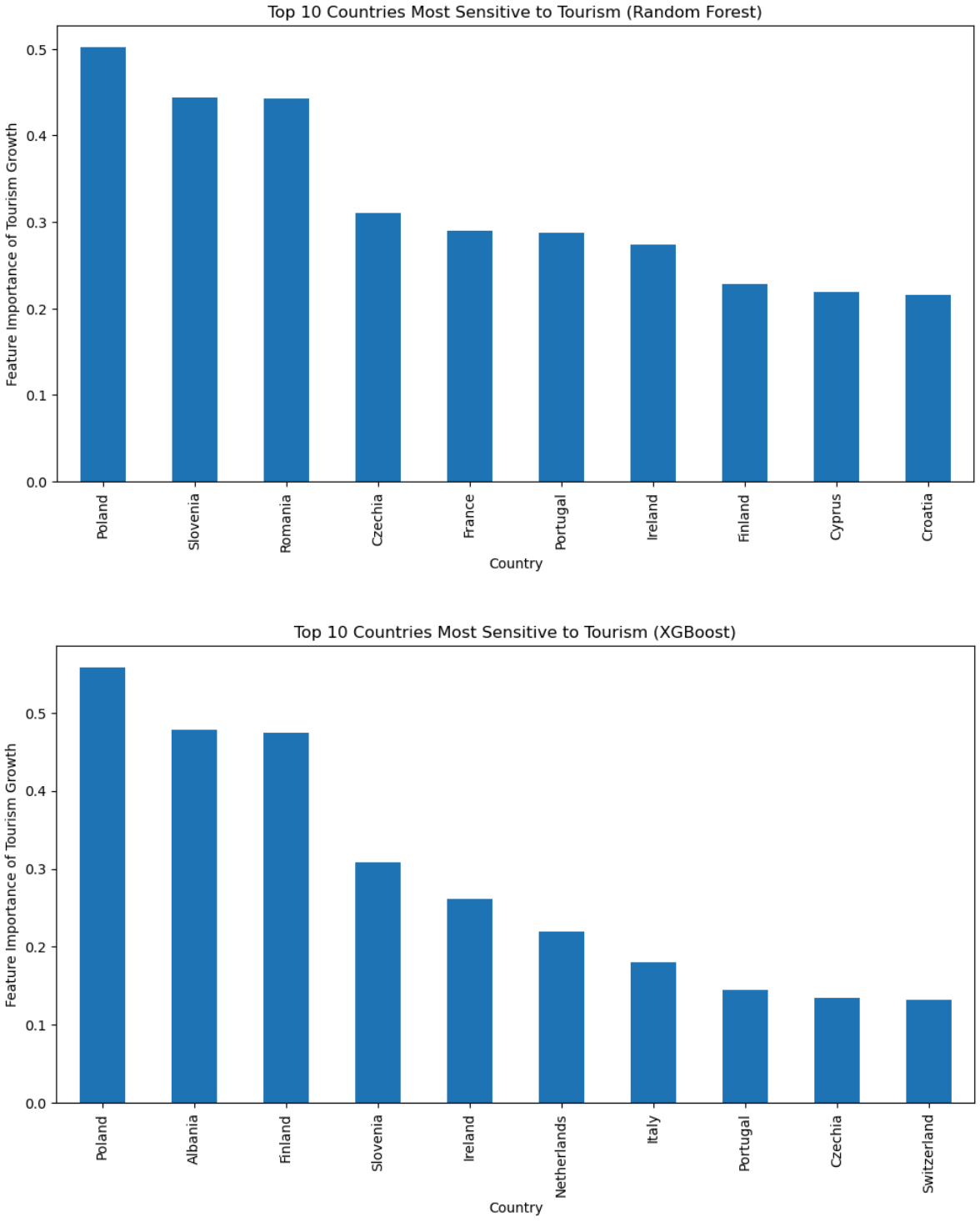

Figure 9: Countries whose GDP growth is most decided by GDP for Random Forest and XGBoost

Conclusions and Discussion

Tourism is indicated to have a heavy correlation with a country's GDP growth. Through our investigations we can find that tourism indicators typically play a strong role in predicting GDP growth contemporarily with our models across the aggregated european country data performing particularly well. These models are able to explain roughly 40% of the variance in GDP through tourism indicators and some control variables. This goes to prove the importance of tourism on a continent wide basis for European GDP. Despite this, while the model generalizes better than per-country forecasts due to the larger dataset, it should not be interpreted as a strong predictive model for year-ahead GDP, as temporal dynamics and country-specific effects are not explicitly accounted for.

In regards to the per country modeling, tourism is shown to have a contemporaneous relationship with GDP growth. The feature importances of the models indicate which countries' GDP is most sensitive to tourism activity, with tourism-heavy economies (e.g., Slovenia, Poland, Finland) showing the strongest relationships. The time series cross validation reveals limited predictive power, and that tourism growth alone cannot reliably forecast future GDP growth, with negative or low R^2 values highlighting high volatility and the influence of other macroeconomic factors.

These models show that although tourism is not a perfect predictor of European economies, it has a strong relationship with the GDP growth of European countries and thus it is important to look at especially for sensitive countries like Finland. If this investigation had been done again I would have likely done more data preprocessing to allow the use of other model types than trees to see if we could get better performance out of them. This would come with a drawback of no explicit feature importances, ultimately making the second investigation harder. Additionally, to improve this investigation, I would aggregate more data from other sources such as Eurostat or UNTWO (World Tourism Organization) to try and get more features to evaluate and more data. With this, I think investigating further into the country specific modeling

scenarios for countries that are extremely sensitive to tourism would prove to be valuable to further figure out the relationship between tourism and GDP in these extreme cases. However, this project still indicates that tourism is a good indicator of a country's GDP and informs upon the relationships between GDP and tourism for both Europe as a whole and on the country level.

Work Cited

Eurostat. *Europa.eu*, 2022, ec.europa.eu/eurostat/databrowser/.

“Tourism Statistics Database.” *Untourism.int*, 2024, www.untourism.int/tourism-statistics/tourism-statistics-database.

World Bank. “World Bank Open Data.” *World Bank*, 2025, data.worldbank.org/.