London Tourism Analysis

by Shuwei Deng, Elvis Leng, Mengqi Li, Jeevanandhan Ramamoorthy

Master of Science in Business Analytics, December 2018,

George Washington University School of Business

A Thesis submitted to

The Faculty of

The School of Business

of The George Washington University in partial fulfillment of the requirements for the degree of Master of Science in Business Analytics

December 12th, 2018

Thesis directed by

Dr. Shivraj Kanungo

Chair of the Department of Decision Sciences

Associate Professor of Decision Sciences & of Info Systems & Tech Management

Acknowledgements

We would like to thank Deloitte for providing the topic and for mentoring our team throughout this project.

We would also like to thank Dr. Shivraj Kanungo and Dr. Larry Yu of the George Washington University School of Business for guiding our team on the clarification of the variables in the dataset used in this project.

We would also like to thank professor Patrick Hall for offering his support for this project.

Abstract

London Tourism Analytics

Tourism is one of the major revenue contributors in London. The city of London's local ministry is interested in knowing how tourism affects London's economy and the behaviors of London's tourists. The ministry hired Travel The World management consulting agency to conduct the research for them. This project was performed to cater to the needs of the London government to understand the driving factors of people overstaying their approved visa period based on their identified nationalities. This prediction could help the local government and the national security agencies to arrive at a course of action to ensure that the citizens of London have an increased safety and a balanced socio-economic well being generated from tourists. We analyzed data sets including as tourists' previous travel history in combination with the prosperity indicators of the country from where the tourists arrive. The team found that the number of visits the tourist had prior to the current visit, the amount of dollars spent during their previous visits, the purpose of their visits, and the region they come from largely influenced the number of days they are going to stay in London.

Executive Summary

This project was performed in order to understand the spending and visiting patterns of the London's tourists from different countries, and what characteristics will lead them to overstay their approved visa period. The Greater London tourism authorities, the MI5 and the Prime minister of London are concerned about the safety and the socio-economic well-being of their citizens which could be largely affected by the tourists who overstay their approved visa period.

The City of London's local ministry is seeking advisory services regarding the following issues:

1. What are the spending and visiting patterns of London's tourists?
2. The difference between EU nations and non-EU nations on their spending and visiting patterns in London.
3. What are the main factors that contributed to the tourists overstay?
4. Which tourists' country of origins have the highest probabilities of overstay?

To address all the concerns, we decided to build a predictive model to find the solutions to the above questions. We used R and Tableau to do the descriptive analysis to understand and explore the given datasets. We employed Python to clean the datasets through various feature engineering techniques and H2O to build models and measure the performance of each model. The team found that the Gradient Boosting Machine (GBM) learning model with a balanced target variable performed the best with an AUC of 81.81%.

In sum, we can now help the London Security agencies to predict whether a person is going to overstay their approved visa based on the details of their country of origin and travel history. Visits, spend, purpose, region, quarter, mode, and GDP are the major driving variables, which greatly influenced the probability of tourists to overstay their granted visa period. Visitors from the countries including Kuwait, Pakistan, and Other Africa have the highest probability of overstaying their granted visa period.

Table of Contents

List of Figures

Glossary of Terms

Market	Countries
Dur_stay	Length of the visit: 1. 1-3 nights 2. 4-7 nights 3. 8-14 nights 4. 15+nights
Mode	Main method of travel: 1. Air 2. Sea 3. Tunnel
Purpose	Main purpose of visit: 1. Holiday 2. Business 3. Study 4. VFR (visit friends or relatives) 5. Miscellaneous
Visits	shows how many visits are represented by a record
Spend	shows the total expenditure made abroad (for UK residents) or in the UK (for overseas residents) during the visit.
Nights	relates to the total number of nights spent whilst on a visit
Sample	is the number of contacts from the main IPS used to support each row of information in the dataset. This can be used as an

	indication of the reliability of the data being examined.
Economic Quality	This pillar ranks countries on the openness of their economy, macroeconomic indicators, foundations for growth, economic opportunity and financial sector efficiency.
Business Environment	This pillar measures a country's entrepreneurial environment, its business infrastructure, barriers to innovation and labor market flexibility.
Governance	This pillar measures a country's performance in three areas: effective governance, democracy and political participation and rule of law
Personal Freedom	This pillar measures national progress towards basic legal rights, individual liberties and social tolerance.
Social Capital	This pillar measures the strength of personal relationships, social network support, social norms and civic participation in a country.
Safety & Security	This pillar ranks countries based on national security and personal safety.
Education	This pillar ranks countries on access to education, quality of education and human capital.
Health	This pillar measures a country's performance in three areas: basic physical and mental health, health infrastructure and preventative care.

Natural Environment	This pillar measures a country's performance in three areas: the quality of the natural environment, environmental pressures and preservation efforts.
GDP	Gross Domestic Product (GDP) is a monetary measure of the market value of all the final goods and services produced in a period, often annually or quarterly. Nominal GDP estimates are commonly used to determine the economic performance of a whole country or region, and to make international comparisons.
EPI	Export Quality Index, measures use export prices as a proxy for quality of exports. Higher values indicate greater quality.
Region	Category of nations under one common groups. For example, Pakistan, India, Srilanka, Nepal, Tibet, Bhutan, Myanmar etc. are grouped under one region called south Asia.
Region2	All Asian countries like India, China, Russia, are all grouped under one larger region called Asia.
ROC AUC	ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes.
False Positive	A false positive error, or in short, a false positive is a result that indicates a given condition exists, when it does not.
False Negative	False negative: A result that appears negative when it should not.

	An example of a false negative would be if a test designed to detect tourists overstay returns a negative result, but the person does have overstayed.
Precision	Precision is a description of random errors, a measure of statistical variability.
Specificity	The proportion of negatives in a binary classification test which are correctly identified.

Chapter 1: Introduction

London Tourism Analytics

The City of London's local ministry has requested consulting service to have a deep understanding of the behavior of its tourists. Over the past years, there is a growing number of tourists visiting London, which brings a great economic benefit to the city. However, this benefit also has some drawbacks including some tourists overstay their granted visa period. Having people overstay their granted visa period will negatively influence the economy of London, and also is a national security issue. The goal of this project is to understand which countries bring the most economic benefits to London, and which countries' tourists have high probabilities overstaying their visa.

This project analyzed the general trend of people's spending pattern in London and did comparisons of EU nations and non-EU nations. The information about people's visa status and whether if a person overstays on their visa are confidential, which we cannot obtain. However, we did find a high correlation between people who stay over 15 days and people overstay their visa. Hence, instead of doing research on who is going overstaying their visa, this project focused on predicting and analyzing tourists from what countries are going to stay over 15 days based on their travel history.

This project is going to answer the following questions:

1. How did the number of visits change for the past years? We mainly used different visualization to demonstrate this change.

2. How did the spending patterns of London's tourists change throughout the years? We also plotted different graphs to visually represented those patterns.
3. For all the other questions regarding what factors that affected the tourists overstay, we built a predictive model and used the model to find the important variables.

Chapter 2: Background

Deloitte consulting

This project was advised by the mentors from Deloitte Consulting. Deloitte Touche Tohmatsu, Deloitte, is one of the biggest consulting and accounting companies in the world. Deloitte also has the highest standards for the work it produces on their consulting projects. Deloitte is interested in helping TTW Tourism Analytics to serve the City of London's local ministry to find what characteristics will contribute to tourists overstay.

Availability of Data

All the granted datasets for this project are open to the general public. The "international-visitors-london-raw.csv" data set is a wrangled data from Office for National Statistics (ONS). All data are taken from the International Passenger Survey (IPS) showing the purpose, duration of stay, mode of the tourists, as well as weighted night, visits, and spend updated by quarter. The other dataset "PI_All_data_2007-2017.xlsx" is a dataset published by "The Legatum Institute Foundation" in 2017. The dataset describes 9 pillars ranking of prosperity, which are economic quality, business environment, governance, personal freedom, social capital, safe & security, education, health, and natural environment.

In addition to the granted dataset, we also used the "GDP" data set which we obtained from the World Bank Group to help with the predictive model. It contains each country's annual GDP. GDP was a very important variable in the final predictive model we built. All of the datasets can be downloaded from the web with public access.

Chapter 3: Description of Work Undertaken

Exploring Datasets

Data set 1: "international-visitors-london-raw.csv"

This dataset contains 11 variables including "Year", "Quarter", "Market", "Dur_stay", "Mode", "Purpose", "Area", "Visits", "Spend", "Nights", and "Sample". This is an aggregated dataset which grouped tourists by country. This is the master data set that we started working upon. Our intention is to find the nationalities of the overstaying tourists, but we don't have a column that states if that tourist has overstayed or not. In order to proceed, we classified the tourists who stay for 15 days and more (15+ days) to be considered overstay because of the high correlation between the people who stay over 15 days and the people who overstay.

While exploring the dataset, we had a hard time finding the interpretation of the variable sample and its importance in our dataset. With the help of our professors, Yu and Kanungo, we finally understood the meaning of the variable "Sample", which is the sample weight of each row in our dataset. We also reached out to the collectors of the dataset, Office for National Statistics (ONS) and found out the data was already averaged by a confidential sample weight. Hence, this sample column is not helpful in the dataset. We excluded the sample column from further analysis.

Total Money Spent and Number of Visitors:

We first looked at the money spent and the number of visitors in London from the year 2002 to 2017. The graph below shows the trend in the number of visitors. The color pattern

demonstrates the change in the money spent. The lighter the blue or the bigger the circles, the more money spent in a specific year.

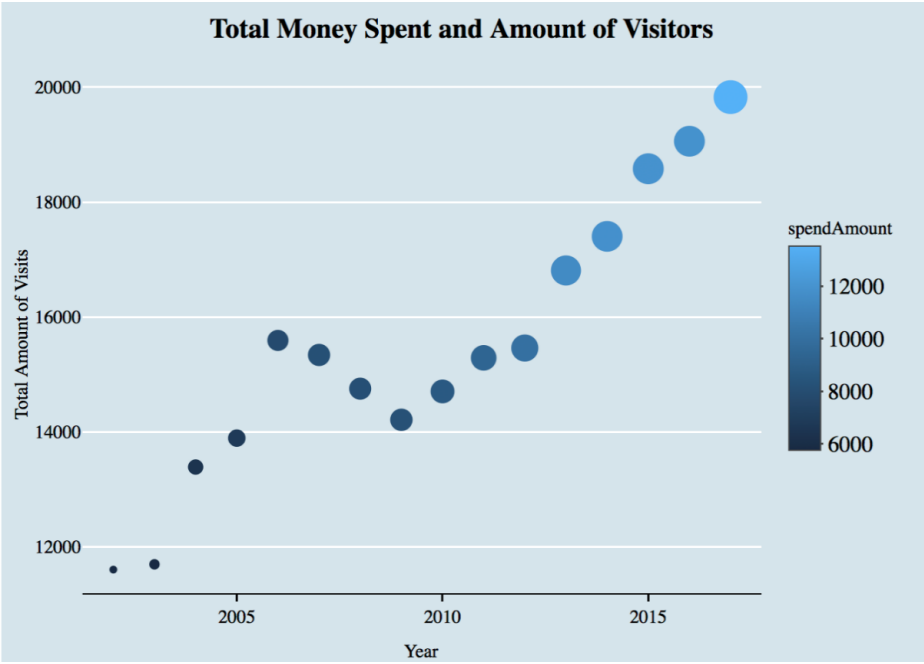

Figure 1. Total Money Spent and Number of Visitors.

By looking at Figure 1, we found that over the period between 2002 and 2017 the number of visitors and the amount of dollar spend both increased across the years. The decline in the number of Visits and Spending starting in the year 2007 is observed. We hypothesized that the decline in the trend is due to the global economic crisis. However, we did not have sufficient information to prove our point at the moment.

Number of Tourists from Different Countries:

We analyzed the number of tourists from different countries using the plot below. The bigger the box or the darker the blue, the more amount of visitors from each specific country.

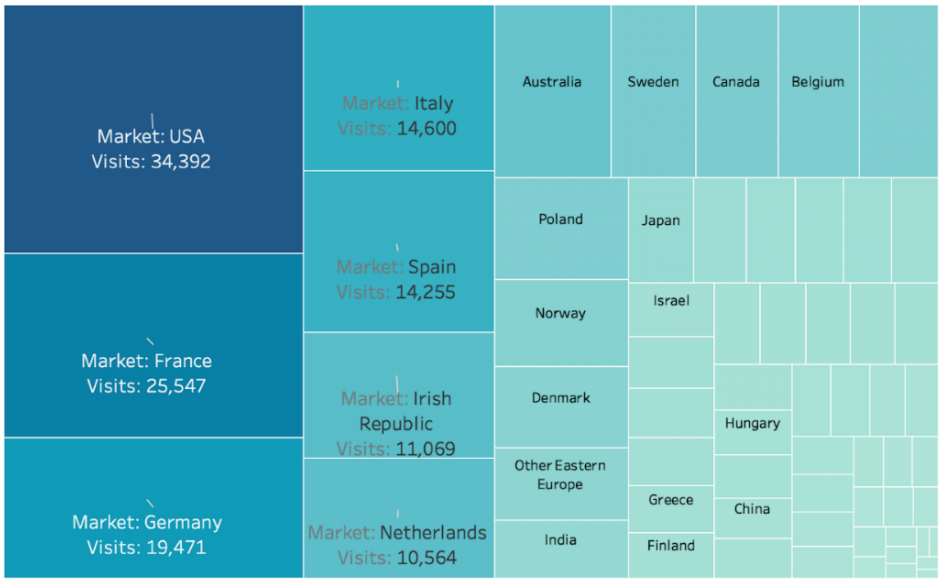

Figure 2. Number of Tourists from Different Countries.

Based on the graph above, we find United States has the largest portion comparing with other countries followed by EU nation and Australia.

Total Spending of Different Countries:

Then, we looked into the countries with the highest amount of spendings. The graph below presented the amount of money each spent in London by different countries. The bigger the box or the darker the blue, the more money each specific country spent.

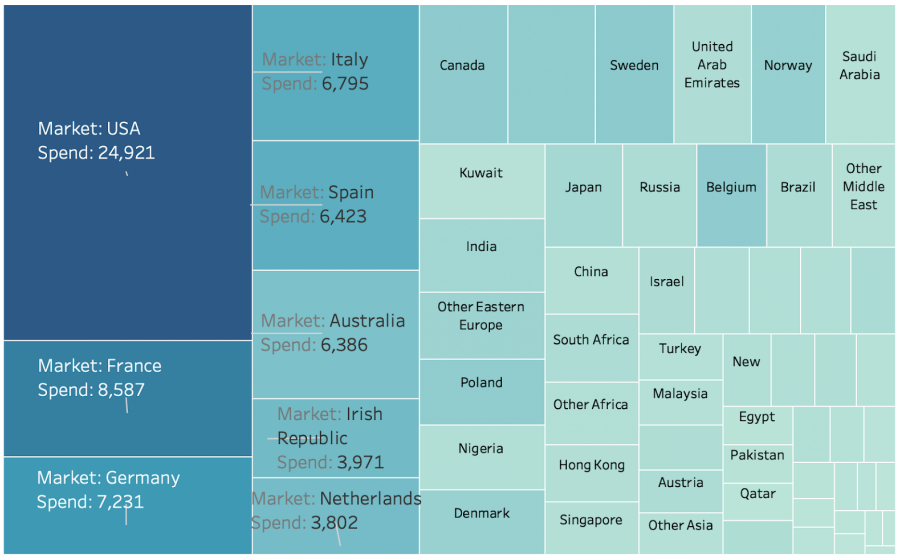

Figure 3. Total Spending of Different Countries.

As we can see from Figure 3, USA spends the most amount of money followed by France and Germany. The money spent by tourists from the USA is almost triple the money spent on tourists from France. As we will see earlier in the report, USA also has the highest number of visitors, which explains why people from the USA spends the most amount of money.

Interestingly, in Figure 3, USA, France, Germany, Italy, and Spain are the top 5 countries that have the greatest number of visitors, which are the same countries with the same orders that presented in Figure 2. In other words, the top 5 countries that have the biggest number of visits also have the corresponding spending power.

Total Visits of EU vs. Non-EU:

We then compared the visiting patterns between the number of visits from EU nations and the number of visits from Non-EU nations. There are total 64 countries in our dataset, including 19 EU countries and 45 Non-EU countries.

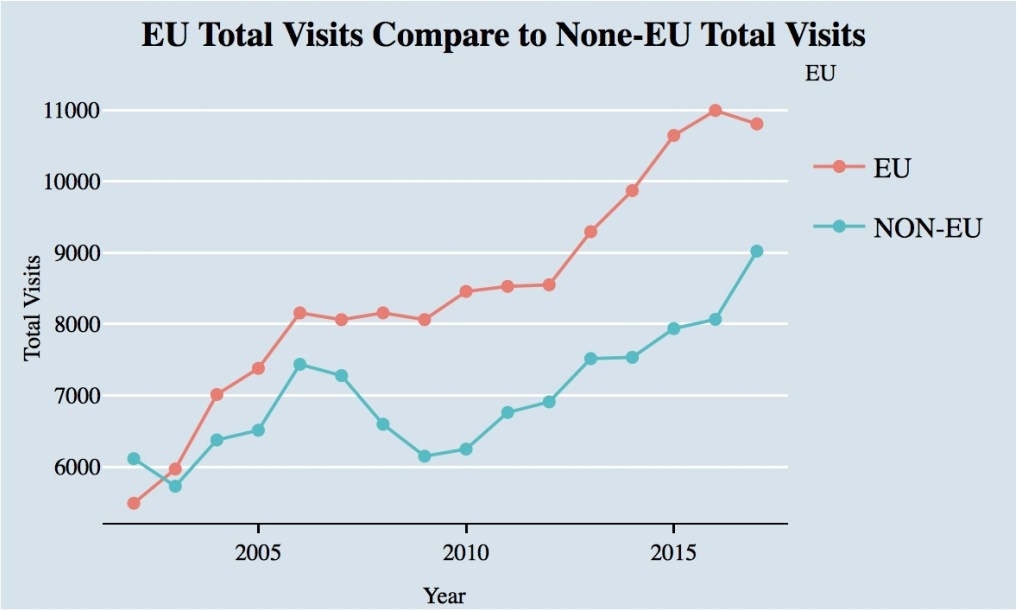

Figure 4. Total Visits of EU vs. Non-EU.

According to Figure 4, the number of visits from Non-EU is almost always more than the number of visits from the EU despite the fact that tourists from the US contribute the most to the number of visits. The number of visits from both EU and Non-EU countries has an overall increase over the year 2002 to 2017. The team observed a dramatic decrease from 2006 to 2009 for the number of visits from Non-EU nations, we assumed it was because of the global financial crisis during that period. In our previous findings, the United States has the greatest number of travelers to London. Moreover, during the financial crisis from 2007 to 2008, there were about 8.8 million people get laid off in the United States alone, while the influence was worldwide (Online, 2018). Since such a great amount of people were impacted during the great depression periods, it was reasonable that people postpone their travel plans to minimum spending and try to save some money in order to survive the hard time, which supported our hypothesis that the financial crisis was the main reason that makes the decrease of number of visits during 2006 to 2009. EU nations were affected less

comparing to the United States, so during that period, the number of visits from EU kept horizontal.

Total Visits by Different Duration:

We wanted to compare the trends of various durations of stays across the years to get a better understanding of how different duration of stays changes.

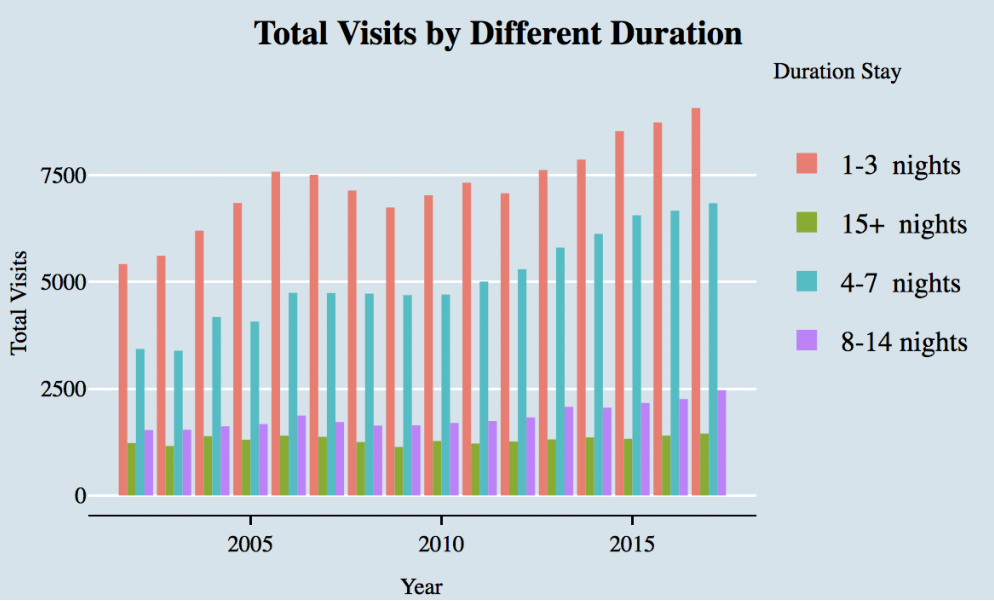

Figure 5. Total Visits by Different Duration.

It can be noted that the 1-3 nights stay and 4-7 nights stay contributed the most of the total stay each year, while 8-14 nights stay and 15+ nights stay contributes the least the total stay each year. In the category that travelers will stay over 8 nights, the total visits appear to be relatively steady over the years. However, people stay 1 to 7 days appear to have a very clear pattern of their stay.

Correlation:

The heatmap beneath indicates the correlation between all variables of the wrangled "international-visitors-london-raw.csv" in order to find the factors that may affect the duration of stay. The team observed strong positive correlations between "spend" and "visits", "spend" and "nights", as well as "spend" and "nights", among which "spend" and "visits" has the highest positive correlation. "Duration of stay" and "visits" have the strongest negative correlation.

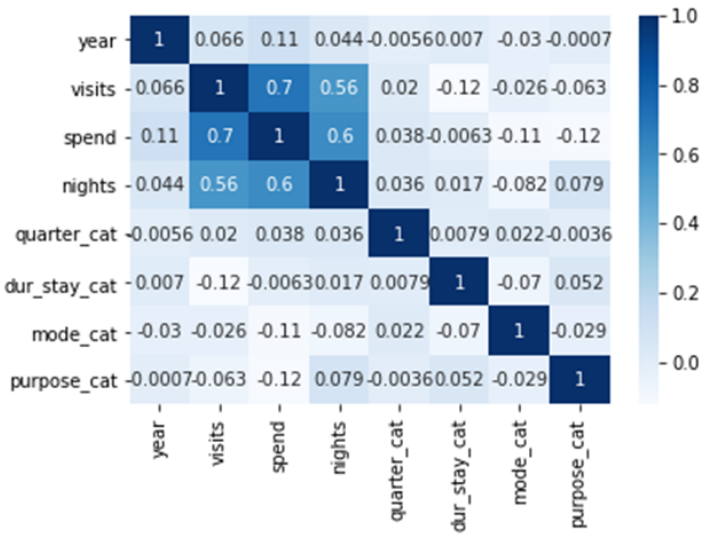

Figure 6. Correlation plot of the variables.

Data set 2: "PI_All_data_2007-2017.xlsx"

The dataset has 26 variables, "country", "isonum", "isocode", "year", "rank_PI", "rank_econ", "rank_busi", "rank_gove", "rank_educ", "rank_heal", "rank_safe", "rank_pers", "rank_soci", "rank_envi", "PI", "econ", "busi", "gove", "educ", "heal", "safe", "pers", "soci", "envi", "region", and "region2". The ranks represent the scores for each prosperity contributors. The dataset has details of both the scores and the corresponding ranks according to the countries. The different geographical locations/countries are grouped into different regions, a new column "region2" is created to store the data. We did not have much difficulty in understanding and interpreting this

dataset. There weren't any assumptions made in this dataset. These are real-time data unlike the previous dataset and the figures quoted in this dataset are real.

Additional Dataset: Data set 3: "gdp.xlsx"

This dataset contains 62 variables such as "Country Name", "Country Code", "Indicator Name", "Indicator Code" and "years" from 1960 to 2017. The explanations for all these terms are included in the appendix. We have a few missing values for the GDP values in the initial years. However, it doesn't affect the analysis since we only include records from 2006 to 2017 as we discussed previously. All the variables are straightforward. There weren't any assumptions made in this dataset.

Data cleaning and Feature Engineering

Binning and renaming the countries:

When we tried to combine the three different datasets based on "country" and "year", there are some differences in how the different countries are named in the datasets. The visitor's dataset denotes "The United States of America" as "United States" whereas the prosperity dataset denotes it as "USA". We changed the names to "USA". There are other mismatched country names in the two datasets, so we also unified them for the purpose of combining three datasets into one.

After combining all the datasets together, we checked for missing values in our dataset. We found the locations which including "Other Western Europe' and "Other Eastern Europe" had missing values because they did not match with the other dataset. Hence, we decided to impute the missing values using random samples from the existing values. For

instance, based on the data and missing value, we are going to assume when the market equals to other middle east, the region is MENA, the Middle East and North Africa region, after we imputed the missing values from the market, we can see only other southern Africa and other African countries left. We are going to impute them as Sub-Saharan Africa.

After joining the datasets 1 and 2 and fixing the problems mentioned above, we moved onto the third dataset "GDP". We had similar problems in the naming conventions of the countries. We decided to use similar methods to solve this problem as we did previously. For instance, we changed the GDPs' country information into the proper information that can be matched with the visitor's data frame. Then we found that variable "year" in the GDP dataset is not an integer value, so we changed its type as an integer in order to join the GDP dataset with visitors table.

Dropping Some Information:

For the PI dataset, we only have records after 2006 while visitor's dataset has records from 2002 to 2017. We want to make sure that the predictive model can produce the best result, so we decided to use data from 2006 to 2017 to do further analysis. We left joined the visitor's table with the PI table based on the market because the PI dataset has a similar structure with the visitor's dataset. Then we joined the GDP table and all the prosperity index variables to our dataset.

Imputing missing values:

After combining the GDP dataset with the Visitor's dataset, we had some missing values due to the mismatch in the country/region naming. Since we have decided to group the

countries according to regions mentioned in the visitor's dataset, we shall now impute the GDP of the missing countries with the average of the regions GDP. For instance, we made eastern Europe and western Europe's GDP equal to Europe's GDP, Taiwan's GDP to the mean of East Asia's GDP. We used the same technique to the other regions including MENA as well as Latin America and the Caribbean.

Handling the overfitting issues:

We found there are some very strong correlations exist among the variables. These strong correlations were overfitting our model. We decided to remove some of the categories that largely influence our prediction. For instance, students are highly likely to stay for over 15 days. For the purpose of the model, we are going to exclude people who are not here to study.

Instead of using individual countries as one of the predictors, we decided to use the region instead of the country name. Because we have over 60 countries in this dataset, such a high number would result in the model overfit dramatically. The region variable combined the countries into different regions based on their geographic location.

Columns names contain "_dtf" stand for distance to frontier approach to the original columns, which is redundant for our machine learning model. Hence, we have decided to drop those columns contain '_dtf' in its name because taking the standard value losses information on the original value.

Since we are doing the prediction on the perspective of the national security check, we would not know how many nights people stayed in London before they actually landed there. We decided to remove the "nights" variables, so the model does not leak any information which means the model does not "see the future".

The Decision Variable:

Since our basic assumption is that all those people who stayed 15+ days are the ones who would potentially overstay, we are going to make the number of days stayed "Dur_stay" as our dependent variable. Originally, variable "Dur_stay" is an ordinal variable with "1-3 nights", "4-7 nights", "8-14 nights", and "15+ nights". We have to convert the values in that column into binary variables. As an example, "0" indicates that tourist has stayed less than 15 days and "1" means the tourist had stayed for 15+ days, making it easier for developing a prediction model.

Training, Validation and Test datasets:

Our final dataset contains data from the year 2006 to 2017. We considered the data from 2006 to 2014 to be our training dataset, from the year 2015 to 2016 to be our validation dataset, and finally the data corresponding to the year 2017 as our test dataset instead of partitioning the data randomly by a percentage. This way of partitioning is more logical and convincing because we believe the time has a dramatic effect in this dataset. Our final model has 73 columns and 35,915 rows.

Chapter 4: Analyses and Results

Building Model and Measuring Performance

We used the H2O package to build models and then measure our model performance. We loaded our final datasets into an H2O cluster, then we identified our dependent and the independent variables. We have a problem with the nature of the distribution of the variables in the dependent column. When training the data, we found that most of our data belong to a single class "0". There are only 6,000 observations classified as class "1" compared with 35,915 observations in total. It's obviously an imbalanced dataset where the number of records of one class significantly higher than the number of records of other classes. It will lead to machine learning algorithms attempt to predict class "0" with the highest accuracy.

One benefit of H2O is that you can balance classes in one step when defining random grid search parameters. The class distribution will be balanced in the model training process once enabled. H2O will either undersample the majority classes or oversample the minority classes. In our dataset, the number of records of class "1" will have the same number of records of class "0".

Since this is a prediction model with a binary variable, the team used the prediction accuracy of the test set based on the AUC value. We also calculated the confusion matrix to see how we are performing in our prediction, in terms of "False Positives", "False Negatives", "Precision", and "Recall". False Negatives, in this case, mean the model predicts that tourists will not stay over 15 days, but they actually overstayed. From the

perspective of national security, how to minimize the number of tourists who potentially will overstay is worth more researching. Moreover, we want to minimize the False negative rate while not dramatically increase the false positive prediction.

Model Selection

Gradient Boosting Machine Model:

Gradient boosting is a machine learning technique for regression and classification problems, which the model tries to improve the previous decision tree model sequentially.

Model 1: GBM with Balanced Target Variable

When the GBM prediction model was equipped with a balanced target variable, the model gave the following results:

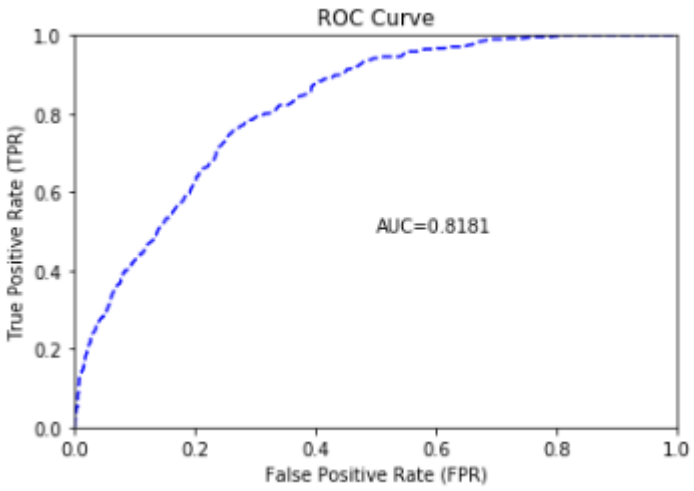

Figure 7. ROC-AUC curve of Model 1: GBM with Balanced Target Variable.

```
ModelMetricsBinomial: gbm
** Reported on test data. **

MSE: 0.10939541087917888
RMSE: 0.3307497707923301
LogLoss: 0.347014209667333
Mean Per-Class Error: 0.25266982463360144
AUC: 0.8180597328570115
Gini: 0.636119465714023
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.164094394029591
5:
```

	0	1	Error	Rate
0	1991.0	691.0	0.2576	(691.0/2682.0)
1	129.0	382.0	0.2524	(129.0/511.0)
Total	2120.0	1073.0	0.2568	(820.0/3193.0)

Figure 8. Confusion Matrix of Model 1.

metric	threshold	value	idx
max f1	0.1640944	0.4823232	255.0
max f2	0.0961476	0.6295366	303.0
max f0point5	0.4865762	0.4652937	97.0
max accuracy	0.5443849	0.8537426	79.0
max precision	0.8708066	0.75	9.0
max recall	0.0151649	1.0	384.0
max specificity	0.9651573	0.9996271	0.0
max absolute_mcc	0.1573410	0.3813232	259.0
max min_per_class_accuracy	0.1657291	0.7436399	254.0
max mean_per_class_accuracy	0.1573410	0.7473302	259.0

```
Gains/Lift Table: Avg response rate: 16.00 %
```

Figure 9. Prediction performance measures of Model 1.

This model has an overall accuracy of 81.81%, which is our best performing model. Since we want to decrease the false negative rate while not dramatically increase the false positive rate. As we can see from the chart above, the precision rate is 0.75 and the recall rate is 1. It is a very balanced rate as recall rate is calculated as the True Positive divided by the sum of the True Positive and False Negative.

Model 2: GBM without Balanced Target Variable

When we built another prediction model without balanced target variable, we found the following results.

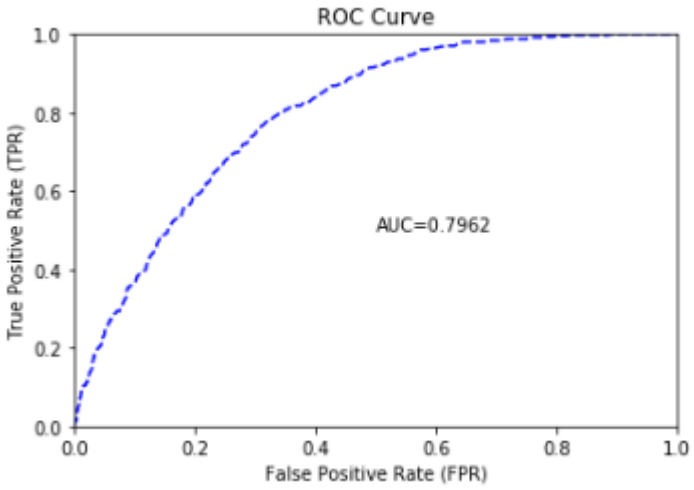

Figure 10. ROC-AUC curve of Model 2: GBM with without balanced Target Variable.

```
ModelMetricsBinomial: gbm
** Reported on test data. **

MSE: 0.11534410267019787
RMSE: 0.33962347190704867
LogLoss: 0.363539466092946
Mean Per-Class Error: 0.270103582482915
AUC: 0.7961911036977691
Gini: 0.5923822073955383
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.157947657930678
4:
```

	0	1	Error	Rate
0	1994.0	688.0	0.2565	(688.0/2682.0)
1	158.0	353.0	0.3092	(158.0/511.0)
Total	2152.0	1041.0	0.265	(846.0/3193.0)

Figure 11. Confusion Matrix of Model 2.

Maximum Metrics: Maximum metrics at their respective thresholds

metric	threshold	value	idx
max f1	0.1579477	0.4548969	234.0
max f2	0.0774718	0.6115203	309.0
max f0point5	0.2825640	0.4130534	145.0
max accuracy	0.5052950	0.8449734	47.0
max precision	0.7362730	0.6666667	2.0
max recall	0.0122491	1.0	387.0
max specificity	0.7661873	0.9996271	0.0
max absolute_mcc	0.1313029	0.3449061	259.0
max min_per_class_accuracy	0.1488385	0.7192394	243.0
max mean_per_class_accuracy	0.1313029	0.7298964	259.0

Gains/Lift Table: Avg response rate: 16.00 %

Figure 12. Prediction performance measures of Model 2.

This model has an overall accuracy of 73.41% which is less than the balanced target predictive model. The precision rate is 0.67, which is also worse than the performance of the previous model. The recall rate is also 1, which is the same as the previous model.

We have also run other general linear models for the prediction. However, the linear models did not perform as well comparing to the tree-based models. Hence, the other linear models are not included in this report. Finally, the GBM with Balanced Target Variable Model performs better than all other models with an accuracy of 81.7% and a False negative rate of 0.285. Thus, we are selecting the Model 1: GBM to build our dashboard and software to solve our case.

Interpretation of results

Question 1: What are the main factors that contribute to the tourists who stay more than 15 days

We can answer this question by finding the variable importance of our driving variables.

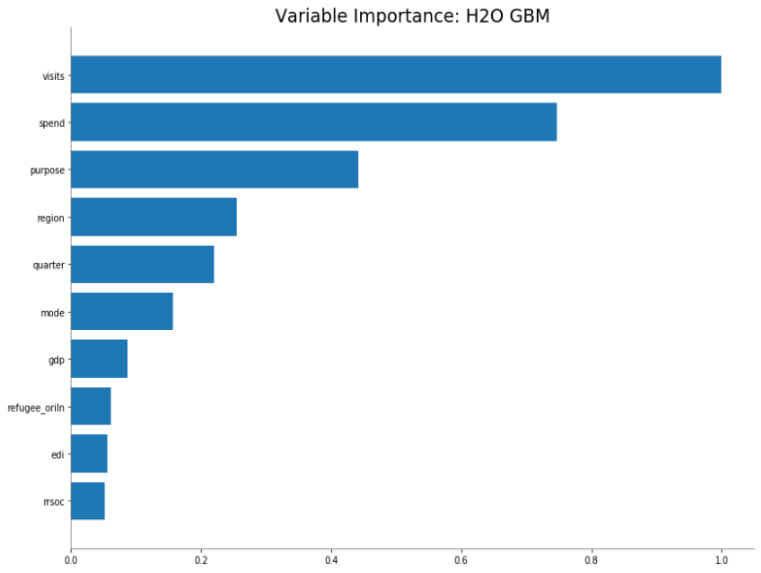

Figure 13. Variable importance plot from the GBM 1 model.

We can see from the above figure that the variable "Visits" tops the list, which means that the number of visits that the tourist had prior to this trip is a critical factor in determining the present duration of the stay. This is followed by the amount of dollars spend during their prior visit, the purpose of their visit, the region from where the tourist comes from, the quarter in which the tourist comes, the mode (air, sea or tunnel) in which the tourist arrives, the GDP of the country that he/she comes from.

The following graph explains how visits affect the prediction of the probability of whether a person is going to overstay their approved visa period. The higher the visit, the less likely this person is going to overstay their approved visa period.

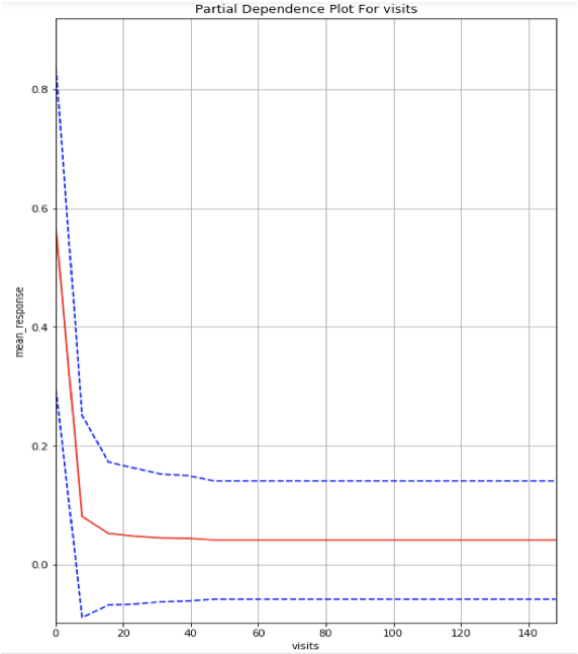

Figure 14. Partial Dependence Plot For Visits.

Question 2: What are the countries that have higher probabilities of staying over 15 days?

Initially while exploring our variables we raised a question that whether the countries that send more tourists have a larger tendency of sending tourists who could possibly overstay. This plot answers that question, USA, France, Australia, and Germany showed a larger tourist ratio, but they are nowhere in the top 10 countries that have higher probabilities of overstaying, with an exception of Australia. We could infer from the prosperity index that the top 10 countries with higher probabilities of overstay are developing countries with a similar socio-economic condition. Also, these countries were once a part of the Commonwealth countries and the UK has strong ties with these nations. This also explains the position of Australia in the top 10 list.

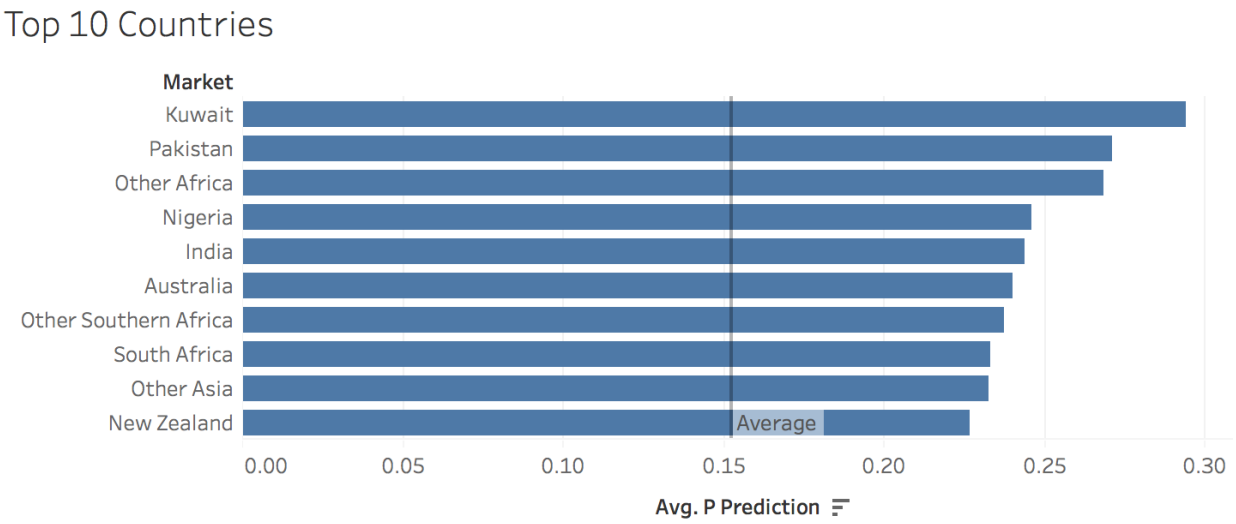

Figure 15. Top 10 Countries with higher probabilities of overstaying.

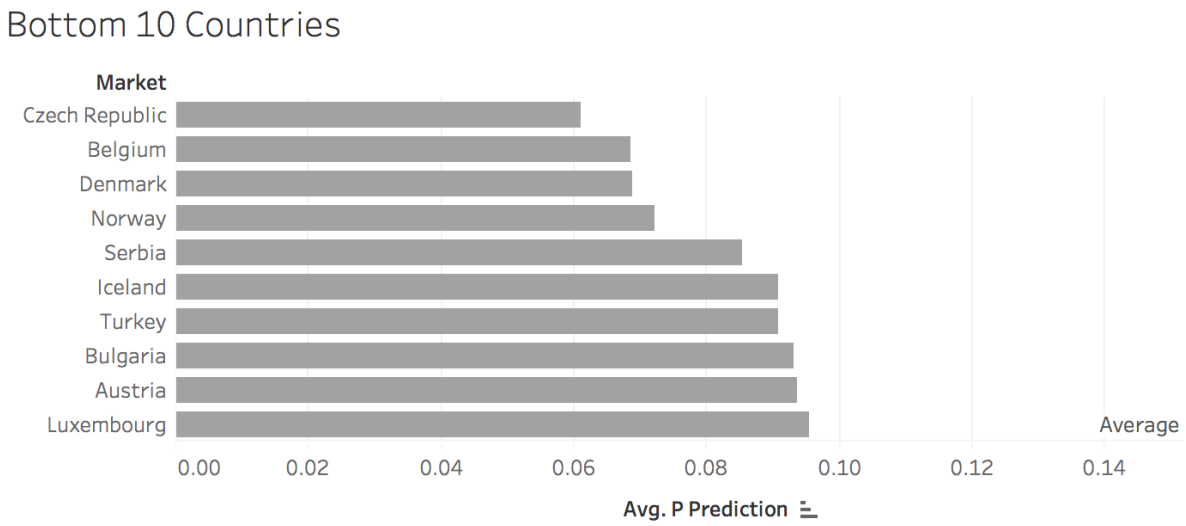

Figure 16. Bottom 10 Countries with higher probabilities of overstaying.

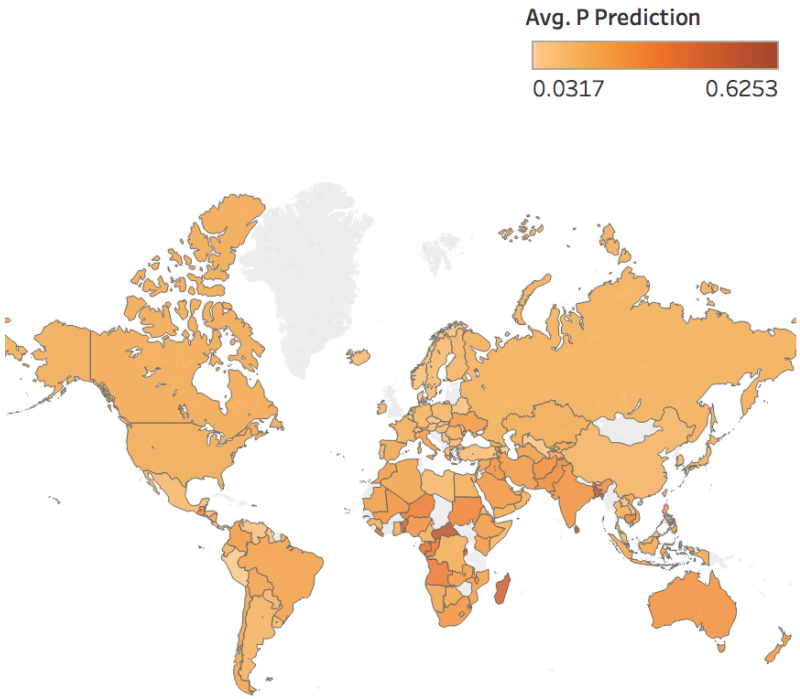

Figure 17. Distribution of countries with the probability of overstaying tourists.

When we analyze the bottom 10 countries or the countries with the least number probabilities of overstaying, we find them mostly to fall into the EU bucket which has a similar socio-economic condition. Thus, it is very evident from these plots that the tourists' country of origin and their country's socio-economic condition does influence the tendency of tourists overstaying.

Question 3: What are the chances of a tourist overstaying the approved visa period?

For those of tourists that stay over 15 days, the potential probability that they will risk the immigration status. This probability can be predicted using our GBM 1 model with an AUC value of 81.81%. To understand the method of prediction let us explore a surrogate tree as given below.

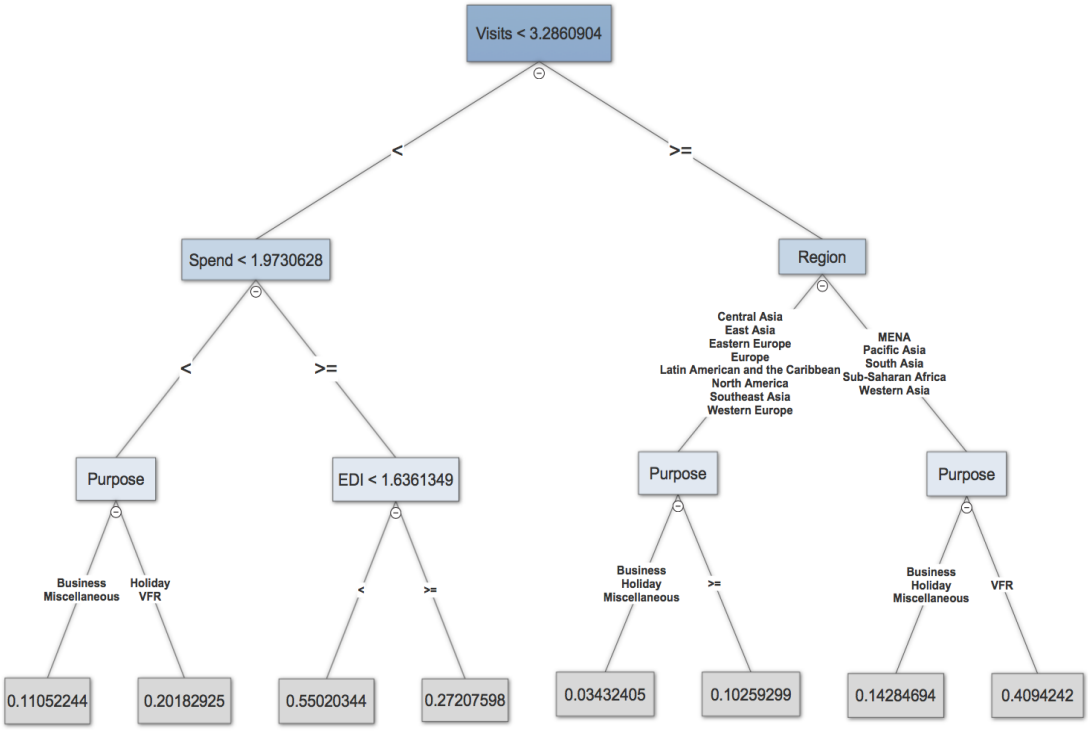

Figure 18. Decision Tree Surrogate Model.

The first split with respect to determining the probability starts with the number of visits. If the tourists had more than or equal to 3.28 visits, then the next split is done based on the region followed by the purpose of visit. If the tourist had less than 3.28 visits previously, then the next split is based on the number of dollars spent during the previous visits, and later by the purpose of visit and the export diversification index. For instance, let us consider a person who visited 2 times on a Business purpose, and spend about $1,500 during his previous visit, let's predict his probability of overstaying. Since he had visited twice (which is less than 3.28) before he would then be evaluated based on his spending, since he had spent only $1,500 < $1,900 he would then be evaluated based on the purpose of visit since he was on a business purpose, the final probability converges to 0.11. So, the probability of this person overstaying is 11%. This is just a depiction of one case of our model. Our model is many times more precise than this simple model. This simple surrogate tree explains the way in which the models are built.

Findings

- Visitors from Non-EU are MORE than visitors from EU.
- The number of visitors from Both EU and Non-EU countries is increasing.
- Visits, Spend, Purpose are the three main contributors that will lead to overstay.

The countries that have high probabilities of overstay are:

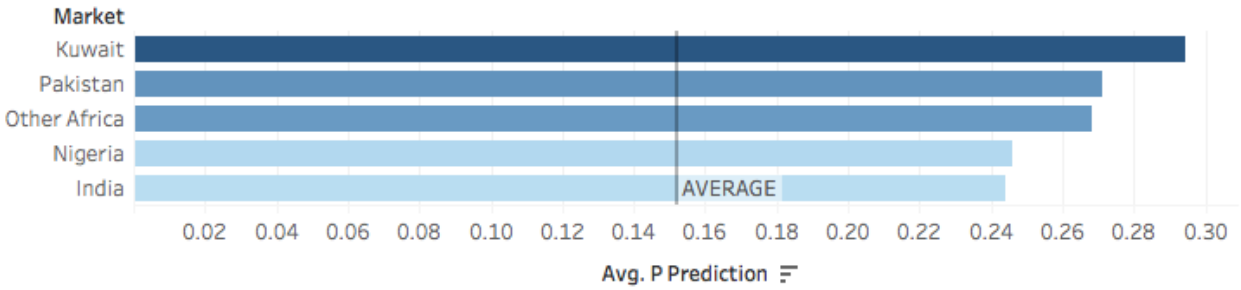

Figure 19. Top 5 Countries with higher probabilities of overstaying.

Chapter 5: Conclusion & Future Improvements

In sum, we have addressed all the concerns of the London Government with respect to their tourism industry and its impact on London's safety and security. We have provided a prediction model, critical variables that help determine the probability of a tourist overstaying the approved during of time and the countries which the highest and the least probabilities of tourist who overstay. This prediction is based on the past data about the tourist travel history to London. We can predict this with 81.81% accuracy and we have also listed that factors such as the number of previous visits, amount spend during the previous visits, the purpose of visits and region of origin are the major factors in determining this probability. The country of origin of the tourists and the country's socio-economic conditions do play a significant role in determining the tendency of a tourist to overstay the allowed duration of stay.

We have done some extensive research into this project, but there is still room for improvements in the future. If we can have data from individual persons instead of weight by countries, we can have a better prediction on the personal level. For the countries that are grouped by regions, if we can have information for those individual countries, we can have a more accurate prediction result for the specific countries.

Chapter 6: Recommendations

From the above results, we are here to make certain recommendations to our clients.

1. Major attention needs to be given to tourists who have visited London less than 3 times
2. Tourists from MENA, Pacific Asia, South Asia, Sub-Saharan Africa, and Western Asia are likely to overstay.
3. Tourists with the purpose of visiting friends and relatives are more likely to overstay.
4. Tourists with lower export diversification index of the country of origin are likely to overstay.
5. Visit: Have restricted rules for Visa application process and tighten the customs inspections for the countries that have a high probability of overstay. Grant different visa durations for different travel purposes.
6. Spend: Limit the money the tourists can bring during their trip.
7. Purpose: Make notes on those tourists who stayed longer than the visa issued time, especially for the tourists that have visits <3.286, next spend < 1.9731, and then EDI < 1.6361.

References

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Legatum Institute Foundation. (2018). *Prosperity rankings. (n.d.).* Retrieved September 1, 2018, from https://www.prosperity.com/about/resources

Office for National Statistics. (2018). *Number of International Visitors to London*. Retrieved September 1, 2018, from https://data.london.gov.uk/dataset/number-international-visitors-london

Online, F. (2018, April 26). How badly 2008 financial crisis hurt US, UK, Germany: This 1 chart shows it all. Retrieved from https://www.financialexpress.com/economy/how-badly-2008-financial-crisis-hurt-us-uk-germany-this-1-chart-shows-it-all/1146819/

The World Bank Group. (2018). *GDP (current US$)*. Retrieved September 2, 2018, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD