IBM AI Enterprise Workflow Capstone

Final Summary Report

Claire Lubash, July 2020

1 Introduction

AAVAIL is a video streaming company whose customer feedback has prompted consideration for an adjustment to their current subscription-based business model. The business objective is as follows:

- Build a tool that, at any point in time, can project revenue for the following month for a specific country.

The data includes transaction-level purchases for 38 countries over a couple years and a few thousand users. A revenue projection tool would save time for managers and provide more accurate predictions, leading to stabilized staffing and budget projections. For ease of use, an API is requested to automate the entire process, including the data ingestion and model deployment.

Testable Hypotheses:

1. An automated prediction tool will be ore accurate than the current method implemented by the managers.
2. The data provided from past years is correlated with future data.

Ideal Data:

- Our goal is to have revenue broken down by country and month. Other features (e.g. country, price, product, etc.) will be needed in order to accurately create our prediction model.

2 Data Analysis

The dataset consists of time series data ranging from October 2017 through July 2019. The transaction data includes revenue, purchases, total views, unique invoices, and unique streams. The data can be broken down by country in order to further analyze revenue trends.

The business is most successful during the final quarter of the year. Revenue, purchases, total views, unique invoices, and unique streams all reach their peak during these months. Peak values are approximately double the average value for the other months.

The United Kingdom is by far the largest market, bringing in just under 90% of all revenue. Over all three years, the United Kingdom is consistently the country bringing in the most revenue. Eight of the top ten countries are located in Europe, with the remaining two from Asia.

	purchases	unique invoices	unique streams	total views	revenue
count	2270	2270	2270	2270	2270
mean	710.52	37.16	383.15	3693.14	3409.97
std	935.07	45.92	424.71	4564.99	8049.97
min	1	1	1	0	0
25%	22	1	22	175	83.62
50%	76	3	72	677	301.06
75%	1329.5	73	780	7032.75	5068.52
max	7756	219	1596	29374	170304.18

3 Visualizations

3.1 Monthly Revenue Data

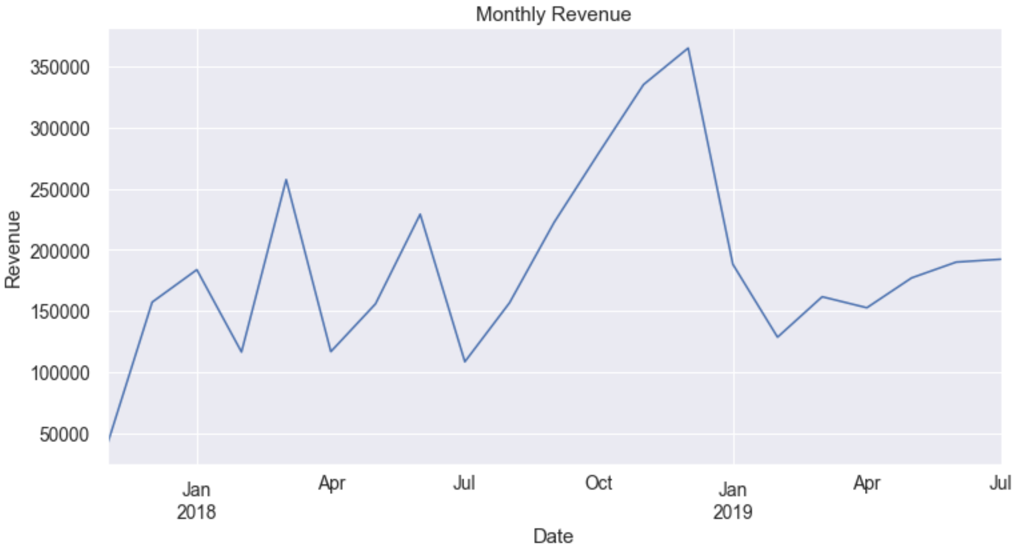

Figure 1: Total monthly revenue for all markets combined.

3.2 Revenue Breakdown by Market

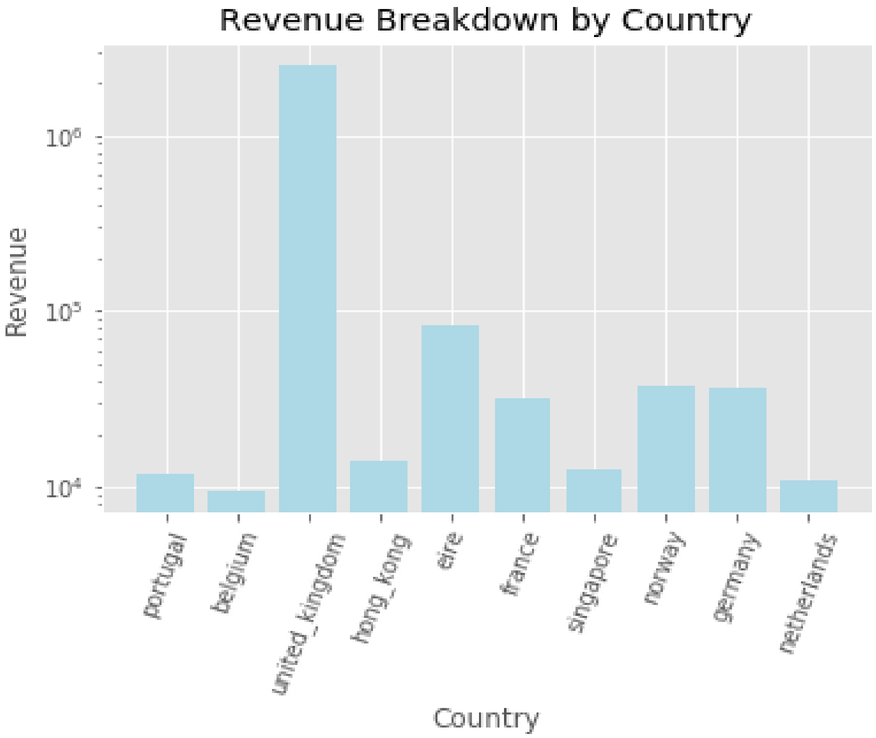

Figure 2: Total revenue brought in by each country.

3.3 Monthly Transactional Data

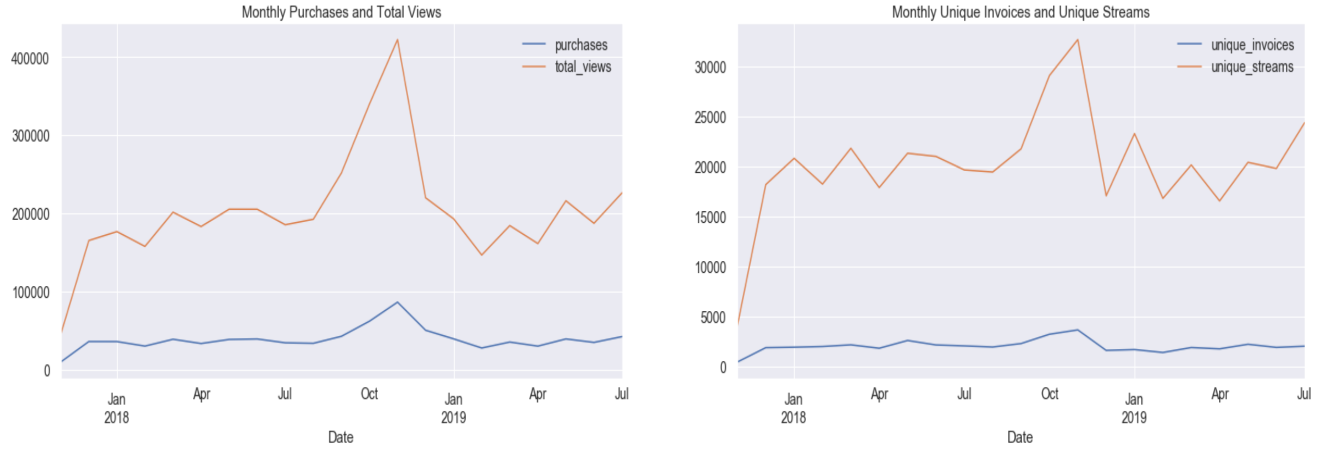

Figure 3: Transactional data for all markets over time.

3.4 Variable Correlation

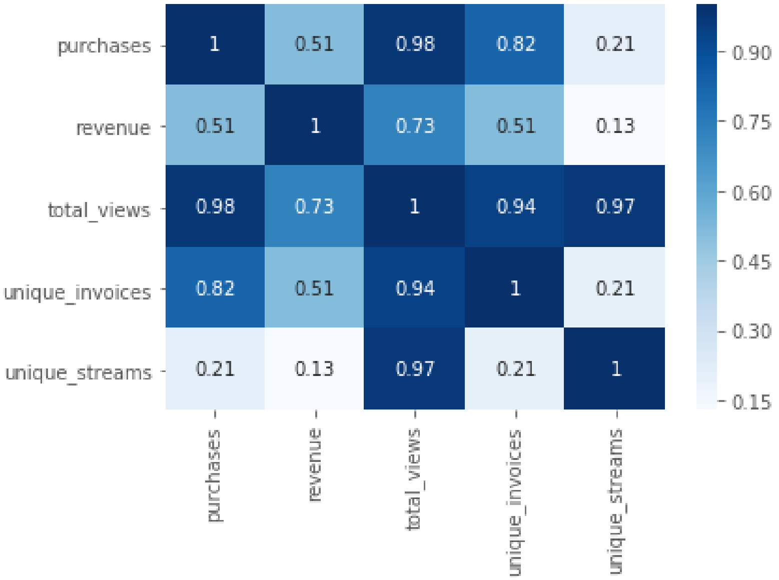

Figure 4: Correlation between the numerical variables from the dataset.

4 Model Selection

We hope to use the different features in our dataset in order to predict a future month's revenue for a given country and date. Because our target value is numeric, we will investigate various regressors and their respective RMSE values for determining our ideal predictive model.

Suite of Models:

- Random Forest Regressor
- Ada Boost Regressor
- Gradient Boosting Regressor

Hyperparameters are tuned for each algorithm with the use of the scikit-learn classes such as `Pipeline` and `GridSearchCV`. This process is repeated for each country's dataset to determine the most appropriate model and parameters for each market.

Adjusted Parameters:

- Number of Estimators
- Learning Rate
- Maximum Number of Features

After iterating through each regressor for each market, it was determined that the Random Forest Regressor was the ideal model for all of the countries. The results are outlined below:

Country	RMSE Values			Optimal Model
	Random Forest	Ada Boost	Gradient Boosting	
All	26533.12	47340.78	38315.53	*RandomForestRegressor*(max_features = 5, n_estimators = 15)
Belgium	98.23	288.45	296.79	*RandomForestRegressor*(max_features = 4, n_estimators = 20)
Eire	2200.40	2836.36	2424.11	*RandomForestRegressor*(max_features = 4, n_estimators = 25)
France	516.01	875.54	796.03	*RandomForestRegressor*(max_features = 4, n_estimators = 20)
Germany	376.38	640.14	626.28	*RandomForestRegressor*(max_features = 4, n_estimators = 25)
Hong Kong	1051.50	1106.03	1102.49	*RandomForestRegressor*(max_features = 3, n_estimators = 25)
Netherlands	95.08	204.55	167.70	*RandomForestRegressor*(max_features = 5, n_estimators = 20)
Norway	233.75	298.08	249.04	*RandomForestRegressor*(max_features = 3, n_estimators = 25)
Portugal	481.45	633.55	680.01	*RandomForestRegressor*(max_features = 5, n_estimators = 20)
Singapore	115.00	210.43	202.42	*RandomForestRegressor*(max_features = 4, n_estimators = 25)
United Kingdom	17626.23	39783.33	32998.50	*RandomForestRegressor*(max_features = 4, n_estimators = 25)