A Data-Driven Approach to Forecasting the US Beer Industry

Team F.Caprasse - A. Scaglia
Faculty R.Mazumder – W.Chen
Company H. Wakil-Moyses

Project Overview

Anheuser-Busch InBev

The **world's largest brewer** (~45% market share in the US) with approximately 500 beer brands and $16B in sales.

Project Importance

- Yearly forecasts of the industry's size are critical to better understand the coming trends, adjust their strategy accordingly and define the company's targets and strategy.
- Historically done manually. Opportunity to use enriched dataset to derive unique insights.

Scope and Objectives

Forecasting
- 2020 U.S. beer market industry forecast
- 2020 Industry forecast for the 1140 designated market areas (DMA)/Segments

Key Factors
- Identify the key factors that influence the growth of the industry
- Quantify the impact of each factor on the industry sales

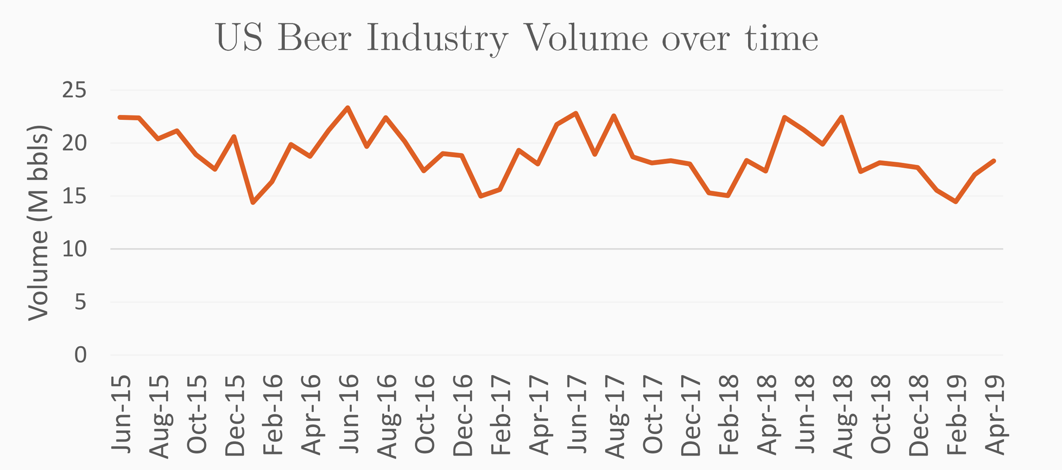
Recommendations
- Optimal resource allocation
- Insightful strategic decisions: how to tweak internal factors for maximum performance

Simulation Tool
- Develop tool to simulate different scenarios
- Maintainability and interpretability

Data

Internal Sources

- **Sales Data** The Beer Institute of Research (BIR). aggregates, anonymizes and redistributes monthly industry data ABI. The data is aggregated by wholesaler and product segment (47 months of data)



- **Pricing Data** Major competitors tend to align their prices on ABI's. Shifts in the overall industry are led by ABI, hence using their pricing data as the industry's is a reasonable business assumption.

External Sources

- **Calendar Data** The number of selling days (weekdays) in a month, important holidays and events.
- **Weather Data** Beer sales are highly seasonal and are influenced by temperature and precipitation that can have a high impact on consumption.
- **Econometric Data** Every DMA has its own context, demographics, history, and preferences (e.g. population).

Challenges and Preprocessing

Business Requirements

- Interpretable models for non-technical employees
- Certain features need to be included (price, weather, etc.)
- Business sense of coefficients and impact of features

Aggregation

DMA/Segment aggregation → 47 monthly data points for 1140 models (June 2015 – April 2019)

Target Generation

Monthly Trend: percentage change of the industry sales volume for each month compared to the same month the previous year → lose 12 months of data

Feature Engineering

- **Ewma6**: 6-month exponential weighted moving average of the trend → lose 6 months of data
- **Price**: Percentage price change
- **Temperature**: Absolute average temperature change
- **Weekdays and Holidays**: Change in number of week-days in the month w.r.t. the previous year

Train, Predict, Validation Procedure

- **Split**: Ordered split 80% - 20% (lose 6 data points)
- **Iterative prediction**: predict one month at a time and recompute ewma6 at each step
- **Validation**: Score the models with cross-validation on both monthly MAE and aggregated MAE
- **Performance metrics**:
  - Monthly MAE → MAE on the monthly trend predictions
  - Aggregated MAE → Error on the trend predictions aggregated for the entire test set (6 months).

Project Timeline

On-Campus Research			On-site Internship		
February/March Literature Review Data Access	April Data Understanding Data Exploration	May Data Preprocessing	June Feature Engineering Modeling – Phase 1 & 2	July Modeling – Phase 3&4 Model Selection	August Tool Building Recommendation

4-phase Modeling

1. Time Series

Seasonal ARIMAs with exogeneous variables

+ Very powerful
+ Works well with little data
+ Automatically deals with seasonality

- Very complex
- Not easily interpretable
- Little control on the parameters

2. Interpretability

Linear regression with hand crafted features

+ Very simple
+ Interpretable
+ Fully flexible

- Medium performance
- Needs more data than available
- Needs to be coded from scratch

3. Generalization

Constrained and regularized regressions

+ Better generalisation
+ Positive coefficients
+ Fully flexible
+ Deals well with high dimensionality

- More complex
- Some coefficients are set to zero
- Needs to be coded from scratch

4. Fine-tuning

Clustered models

Cluster DMAs. Optimize convex combination of individual and clustered model

+ Train on more data
+ Identify trends common to DMAs

- Lose interpretability
- Dependent on clusters
- Needs to be coded from scratch

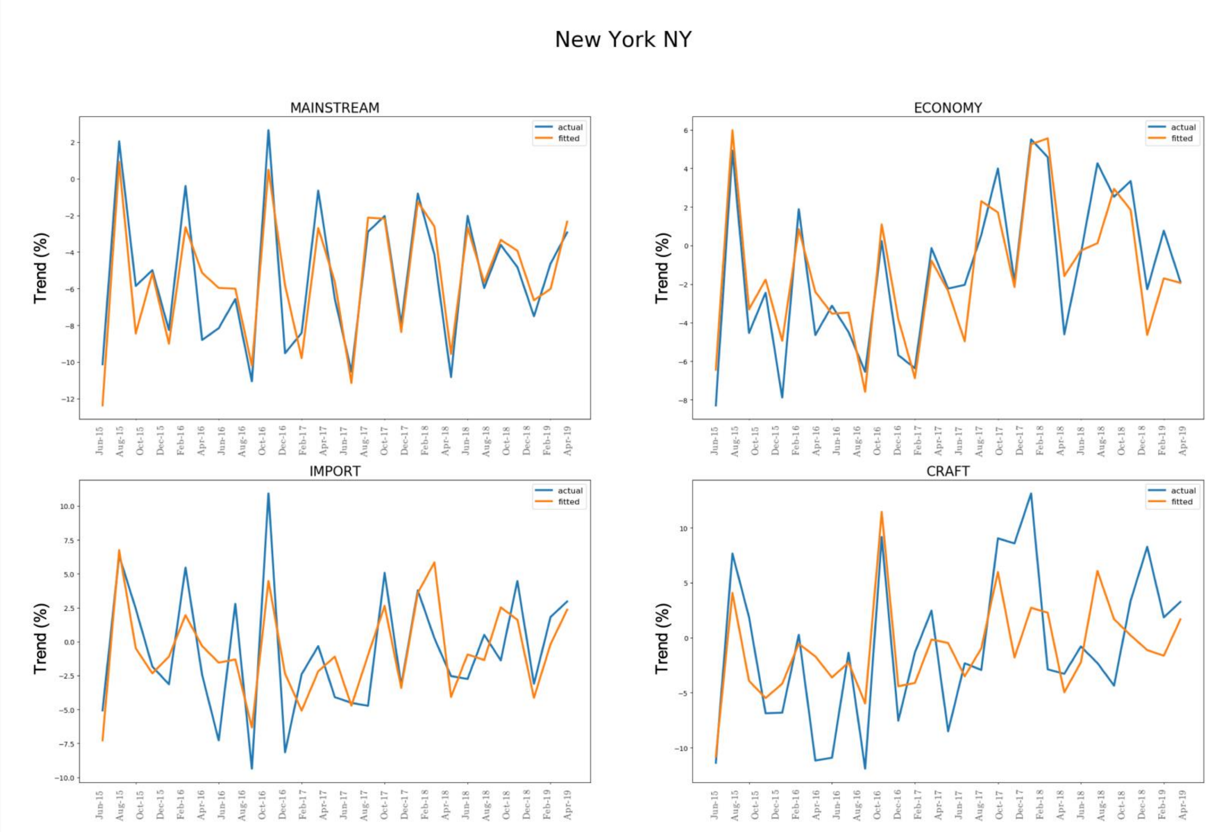
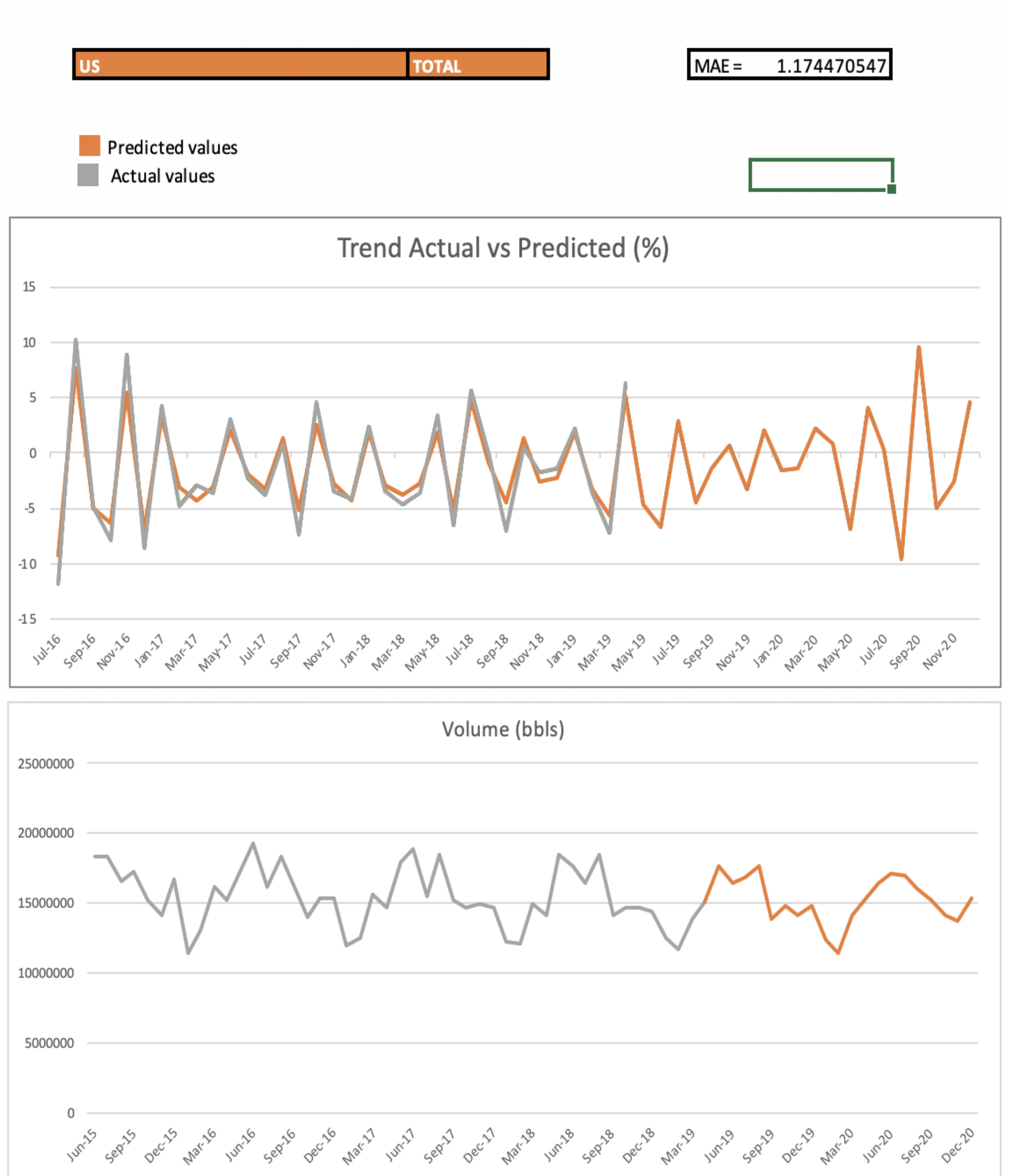


Figure 1. Fitted trend for 4 main segments of the New York DMA

Final Models: 1140 Linear Regression with L1-Norm regularization and non-negative coefficients, no clustering
0.17 MAE on aggregated US trend predictions which exceeds ABI's expectation

Impact: Visualization and simulation Tool

DASHBOARD DMA & SEGMENT PREDICTIONS



Industry Model Yearly Overview

	2016	2017	2018	2019	2020
Drivers - var%					
intercept					
Trend Season		0.00%	0.00%	0.00%	0.00%
Price		0.00%	0.00%	0.00%	1.50%
Average Temperature		0.00	0.00	0.00	0.00
Monday		-0.0203	0.0619	-0.0708	0.0083
Tuesday		-0.0319	-0.0160	0.0555	-0.0843
Wednesday		-0.0101	-0.0210	-0.0113	0.0627
Thursday		0.0111	-0.0051	-0.0139	0.0634
Friday		-0.0398	0.0192	0.0050	-0.0078
Saturday		-0.0488	0.0231	0.0249	-0.0017
Super Bowl		-0.0172	-0.0027	0.0011	0.0050
Fourth of July		0.0000	-0.0179	-0.0065	0.0000
Cinco de Mayo		0.0114	-0.0174	-0.0238	-0.0144
Labor Day		0.0104	0.0179	0.0000	-0.0833
Easter		-0.0235	0.0262	-0.0212	0.0000

Impacts %						
Drivers - var%	**Elasticities**					
intercept	0.00%					
Trend Season	0.00%		0.0%	0.0%	0.0%	0.0%
Price	-44.84%		0.0%	0.0%	0.0%	-0.7%
Average Temperature	0.00%		0.0%	0.0%	0.0%	0.0%
Monday	0.00%		0.0%	0.0%	0.0%	0.0%
Tuesday	0.00%		0.0%	0.0%	0.0%	0.0%
Wednesday	0.00%		0.0%	0.0%	0.0%	0.0%
Thursday	0.00%		0.0%	0.0%	0.0%	0.0%
Friday	0.00%		0.0%	0.0%	0.0%	0.0%
Saturday	0.00%		0.0%	0.0%	0.0%	0.0%
Super Bowl	0.00%		0.0%	0.0%	0.0%	0.0%
Fourth of July	0.00%		0.0%	0.0%	0.0%	0.0%
Cinco de Mayo	0.00%		0.0%	0.0%	0.0%	0.0%
Labor Day	0.00%		0.0%	0.0%	0.0%	0.0%
Easter	0.00%		0.0%	0.0%	0.0%	0.0%
Market Growth forecast			0.0%	0.0%	0.0%	-0.7%
Market Growth (actual/forecast)			-1.7%	-1.4%	-1.6%	-0.7%
Volume (actual / forecast) (000' bbls)		188,447	185,169	182,556	179,564	178,364
Trend Delta from actuals			1.7%	1.4%	1.6%	-0.005%

Next Steps

The results achieved improve significantly on the company's previous status quo. Next steps to explore are:

Implementation

- Knowledge transfer
- Comparison with current methods

Maintainability

- Knowledge transfer to ensure understanding of updating process
- Better software that improves aesthetics and maintainability (e.g. Tableau)

Improvement of algorithm

- More data (weekly or zip code level)
- Predict on a more granular level (e.g wholesaler)
- Advanced methods (e.g. further exploration of clustering and convex optimization)
- Relaxation of business constraints

After refining the model further, we recommending running field experiment with an interesting DMA (e.g. Houston)