# **The Median Group – Assignment 6 part 1**

Ben Haimowitz, Chris Lattanzio, Courtney McCarthy, Jerone Kadner, Natalia Fadeeva

## **Problem Description**

For this assignment, The Median Group is trying to predict the number of minutes viewed for a future time period for one current NBC series on digital streaming platforms (Hulu and NBC app). The group is hoping that by correctly predicting the number of digital minutes consumed for a show, NBC will be able to determine which series have the strongest digital potential in the coming television season.

For future aspects of our capstone, the group hopes to analyze the total viewership of multiple series in the same manner. Total viewership would then be used as a proxy for the success of a show on a digital platform. Once we are able to identify these successful shows we hope to identify other traits that could predict similarity and potentially drive the adoption of new shows in the future. Adoption in this context would be defined as the number of people who view NBC content digitally. NBC would be able to benefit from these insights by hopefully being able to acquire other shows with similar characteristics that are better suited for digital platforms. Then they would be able to market those shows on the Hulu app, the NBC app, and a new over the top (OTT) streaming platform that Comcast (NBC’s parent company) is looking to launch in 2020.

## **The Data**

We will be using a sample of data from digital platforms for the series **This is Us** for its first three seasons, which encompasses 54 episodes that aired from September 20, 2016, through April 2, 2019***.***

## **Defining the Predictive Task**

To understand how well a show does on a digital platform, we have to anticipate the number of minutes viewed of the series on the same digital platform in the future. Furthermore, we would like to understand how many numbers of minutes viewed we could potentially have for the next season of the show.

## **What to predict**

The target variable to predict would be the total number of digital viewership minutes for the series **This is Us**. This will be calculated as the sum of minutes viewed for each of the 54 episodes within the 35 days since an episode became available on digital platforms Hulu and NBC app.

## **Which Predictors to use**

We have to decide how to describe the recent past behaviors of the viewers who watched the show. In order to do this, we need to collect a series of indicators that capture the recent dynamics in the number of viewers. Therefore, we will predominantly use the recent number of viewers for the show on a set timeframe.

We will also create a new dataset aggregating the total number of minutes for 35 days of streaming per episode in a season.

We will use technical indicators defined as the mean and standard deviation that will reflect the numeric summaries or our viewership time series.

## **The Prediction Task and the Data we will use.**

We will use the data from the first two seasons as training data.

We will use the data available from the 3rd season as the testing data.

Our prediction task would be total minutes viewed per episode of This is Us series in the third and fourth quarters of the 3rd season.

## **How predictions will be evaluated**

We will divide the existing data into two-time windows, past observation*s* and future observations*.* We will train the model on the training data set and test the model on the test dataset as outlined above. We will use all three learn-test alternatives (the fixed learning window, the growing window, and the sliding window strategies) for time-series data.

We will then use the Performance Estimation for the Time Series Model with SVM and Multivariate Adaptive Regression Splines (MARS). We will use the Monte Carlo method for our estimation task in order to make sure we do not use any resampling based methods as this would alter the natural order of our data.

## **The Results**

We will run a summary of our performance estimation, plot and visually inspect the results to determine the best model to use.