NIELSEN JOURNAL of MEASUREMENT

FEBRUARY 2017-VOL 1 ISSUE 3

USING MACHINE LEARNING TO PREDICT FUTURE TV RATINGS

By Scott Sereday and Jingsong Cui, Data Science, Nielsen

EDITOR-IN-CHIEF
SAUL ROSENBERG

MANAGING EDITOR
JEROME SAMSON

REVIEW BOARD

PAUL DONATO
EVP, Chief Research Officer
Watch R&D

KATHLEEN MANCINI
SVP, Communications

MAINAK MAZUMDAR
EVP, Chief Research Officer
Watch Data Science

FRANK PIOTROWSKI
EVP, Chief Research Officer
Buy Data Science

ARUN RAMASWAMY
Chief Engineer

ERIC SOLOMON
SVP, Product Leadership

The world of measurement is changing.

Thanks to recent advances in data collection, transfer, storage and analysis, there's never been more data available to research organizations. But 'Big Data' does not guarantee good data, and robust research methodologies are more important than ever.

Measurement Science is at the heart of what we do. Behind every piece of data at Nielsen, behind every insight, there's a world of scientific methods and techniques in constant development. And we're constantly cooperating on ground-breaking initiatives with other scientists and thought-leaders in the industry. All of this work happens under the hood, but it's not any less important. In fact, it's absolutely fundamental in ensuring that the data our clients receive from us is of the utmost quality.

These developments are very exciting to us, and we created the Nielsen Journal of Measurement to share them with you.

WELCOME TO THE NIELSEN JOURNAL OF MEASUREMENT

SAUL ROSENBERG

The Nielsen Journal of Measurement will explore the following topic areas in 2017:

bd

BIG DATA - *Articles in this topic area will explore ways in which Big Data may be used to improve research methods and further our understanding of consumer behavior.*

s

SURVEYS - *Surveys are everywhere these days, but unfortunately science is often an afterthought. Articles in this area highlight how survey research continues to evolve to answer today's demands.*

ns

NEUROSCIENCE - *We now have reliable tools to monitor a consumer's neurological and emotional response to a marketing stimulus. Articles in this area keep you abreast of new developments in this rapidly evolving field.*

a

ANALYTICS - *Analytics are part of every business decision today, and data science is a rich field of exploration and development. Articles in this area showcase new data analysis techniques for measurement.*

p

PANELS - *Panels are the backbone of syndicated measurement solutions around the world today. Articles in this area pertain to all aspects of panel design, management and performance monitoring.*

t

TECHNOLOGY - *New technology is created every day, and some of it is so groundbreaking that it can fundamentally transform our behavior. Articles in this area explore the measurement implications of those new technologies.*

USING MACHINE LEARNING TO PREDICT FUTURE TV RATINGS

BY SCOTT SEREDAY AND JINGSONG CUI *Data Science, Nielsen*

INTRODUCTION

Nielsen's TV ratings have been a mainstay of the U.S. media industry for over half a century. They're used to make programming decisions and have become part of our popular culture[1], but they are also the basis for billions of dollars' worth of advertising transactions every year between marketers and media companies. They help measure the success of TV shows, verify that their audience size and composition are delivering against strict media-buy targets, and provide a basis for make-goods if the numbers come up short. From that point of view, TV ratings are metrics that measure the past, or at best the present, of TV viewing.

But ratings are also used to predict the future. They set expectations and affect programming decisions from one season to the next, and they help set the cost of advertising (advertising rates) well in advance of when a program goes on the air. In the U.S. for instance, TV networks sell the majority of their premium ad inventory for the year at the "upfront," a group of events that occur annually each spring. For each network, the upfront is a coming-out party to introduce new programs and build up excitement for the upcoming season, but behind the curtains, it's very much a marketplace for advertisers to buy commercial time on

[1]See the weekly top-10s here: http://www.nielsen.com/us/en/top10s.html

television well ahead of schedule. Upfronts are effectively a futures market for television programming, and they provide networks with some stability in their financial forecasts.

As a result, media companies have invested considerable effort to project future ratings. Reliable forecasts can help industry players make faster, more accurate and less subjective decisions, not just at the upfront, but also in the scatter planning[2] that occurs during the season. And if reliable forecasts can be produced through an automated system, they can be used to enable advanced targeting on emerging programmatic TV platforms.

But ratings projections are challenging: They require a steady inflow of rich, granular, reliable data, and the ability to adapt and incorporate new data to account for the latest changes in viewing behavior. Viewers are increasingly consuming media on different devices and through different channels. Their viewing is also increasingly likely to be time-shifted to work conveniently around their own schedule. These changes are making predictions more difficult. More difficult, but also more crucial to the evolving TV ecosystem.

In this paper, we discuss a recent pilot project where Nielsen worked with one of our key clients to innovate and improve the practice of ratings projections. Through collaboration, we aimed to develop a more accurate (better performance metrics), more efficient (better cycle time) and more consistent (reduced variability) system to improve their existing practice and lay the foundation for an automated forecasting infrastructure.

CHOOSING THE RIGHT DATA FEATURES

What were the parameters of this project? We were asked to make projections for several TV networks. These projections needed to include live and time-shifted program and commercial viewing for more than 40 demographic segments. They also needed to be supplied for each day of the week and hour of the day. For upfront projections, we were limited to utilizing data through the first quarter of the year (Q1), because of the timing of the upfront, and needed to project ratings for the fourth quarter (Q4) of that year all the way to Q4 of the following year.

In every predictive modeling project, the type and quality of the input data have a very significant impact on the success of the model. We considered several factors during the design stage to choose the most appropriate and effective data for this research. It's important to point out how some data, while promising and completely suitable for other types of research studies, can be inadequate or inefficient for our purpose.

Consider, for example, the enthusiasm that a top executive might have for a new program on the lineup. That enthusiasm is hard to quantify. It introduces bias (the executive might have played a larger role in bringing that program to life), and even if we were able to express it mathematically, we couldn't obtain the same information for all the other programs on the air. Domain expertise, in the form of subjective insights, can be invaluable to help guide the design of a predictive model and validate its results, but it often falls short as a direct input variable.

We also needed to ensure that the data would be available on a timely basis—so that it could be digested and utilized in the execution of any future projections. Obviously, post-hoc data (such as post-premiere fan reviews) can be highly indicative of the enduring success of a program, but since it occurs after the program airs, it's useless for projection purpose.

Finally, in order to develop a process that can scale to handle all channels, programs, and dayparts, we decided to only use data that is already stored and managed with some level of automation in current practice. Future programming schedules, for instance, could most certainly boost the accuracy of our models, but they're not currently standardized nor universally available.

In the end, we decided to rely almost entirely on historical ratings data as input to our forecast model. Fortunately, at Nielsen, we've been collecting top-quality ratings data for decades, with rich, consistent and nationally representative demographics information. We included standard commercial and live ratings data in our input variables, as well as time-shifted viewing, unique viewers (reach), average audiences (AA%), persons or households using TV (PUT/HUT), as well as various deconstructed cuts of data. To supplement the TV ratings, we looked at ratings from Nielsen Social, marketing spend (from Nielsen Ad Intel) and other available program characteristics. Fig. 1 highlights some of the data we evaluated for upfront and scatter predictions:

[2]Scatter Planning refers to a small percentage of ad inventory that is reserved by networks for last-minute use.

FIGURE 1: DATA VARIABLES EVALUATED FOR UPFRONT AND SCATTER PREDICTIONS

	DESCRIPTION	EXAMPLE DATA	RATIONALE
PROGRAM CHARACTERISTICS	Known elements to assess and categorize a show	Genre Air date/time	Differences in characteristics impact ratings
PROGRAM PERFORMANCE	Performance on measurable dimensions	Historic ratings	Past performance indicative of future ratings
PROMOTIONAL SUPPORT	Investment in driving awareness among audience	Marketing spend On/Cross-air promos	Greater promotion / spend lifts ratings
AUDIENCE ENGAGEMENT	Audience interest and commitment to a show	Television Brand Effect	Higher intent to watch/ sustained engagement lifts ratings
SOCIAL/ON-LINE BEHAVIOR	Social media information	Nielsen Social Content Ratings	Inbound social media reflects program popularity and engagement

USING EXPLORATORY ANALYSIS TO GAIN INSIGHTS

It's always a good idea to explore the data before building a model. This preliminary analysis doesn't need to be very sophisticated, but it can be crucial to reveal the rich dynamics of how people have watched TV in the past, and it can help highlight some important and interesting factors that will influence our final projections.

Fig. 2, for example, confirms that among the networks that are part of this project, primetime viewing is still by far the most popular daypart for television consumption. Not surprisingly, weekend usage in the daytime is higher than weekday usage. And over the course of the past five years, the overall percentage of persons watching traditional linear television has been trending downward. Note as well the seasonality of the metric.

In Fig. 3, we can see the differences in usage level by age and gender for those same networks, with older viewers much more likely to watch TV than younger generations, and women in each age group typically watching more than their male counterparts.

FIGURE 2: PERCENTAGE OF PERSONS USING LINEAR TV FROM 2011 TO 2016 (PERSONS 25-54, LIVE+7)

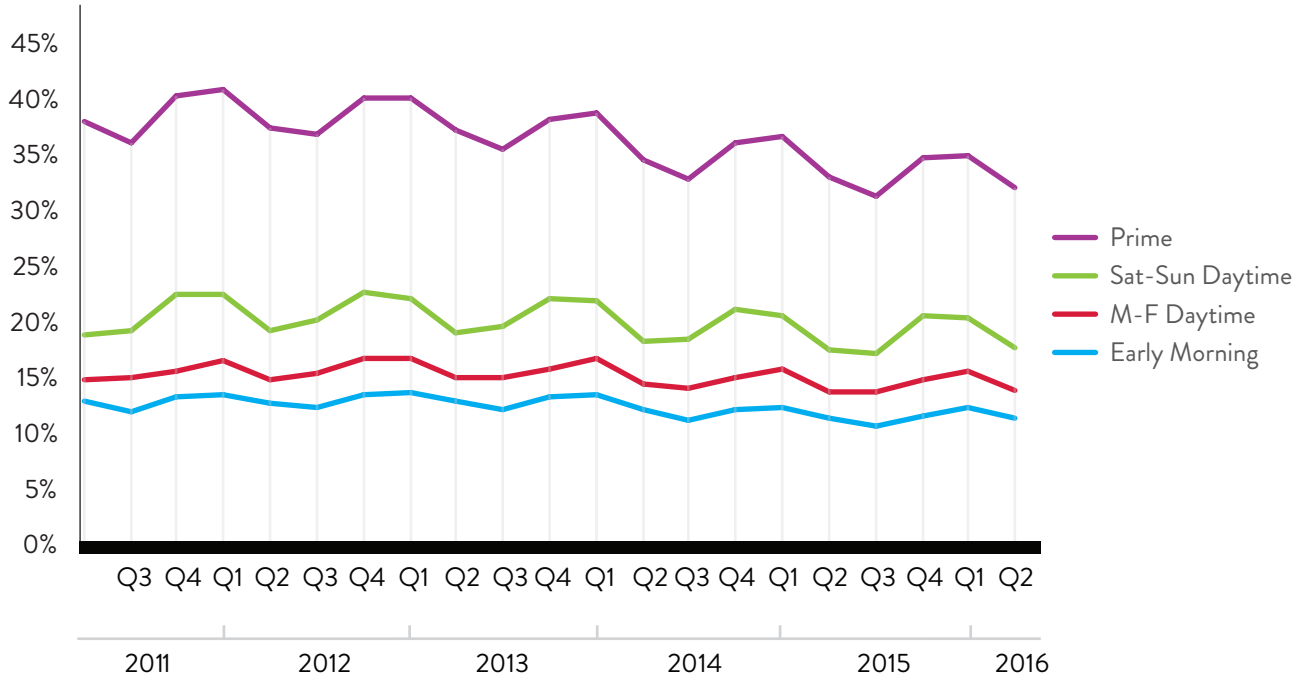
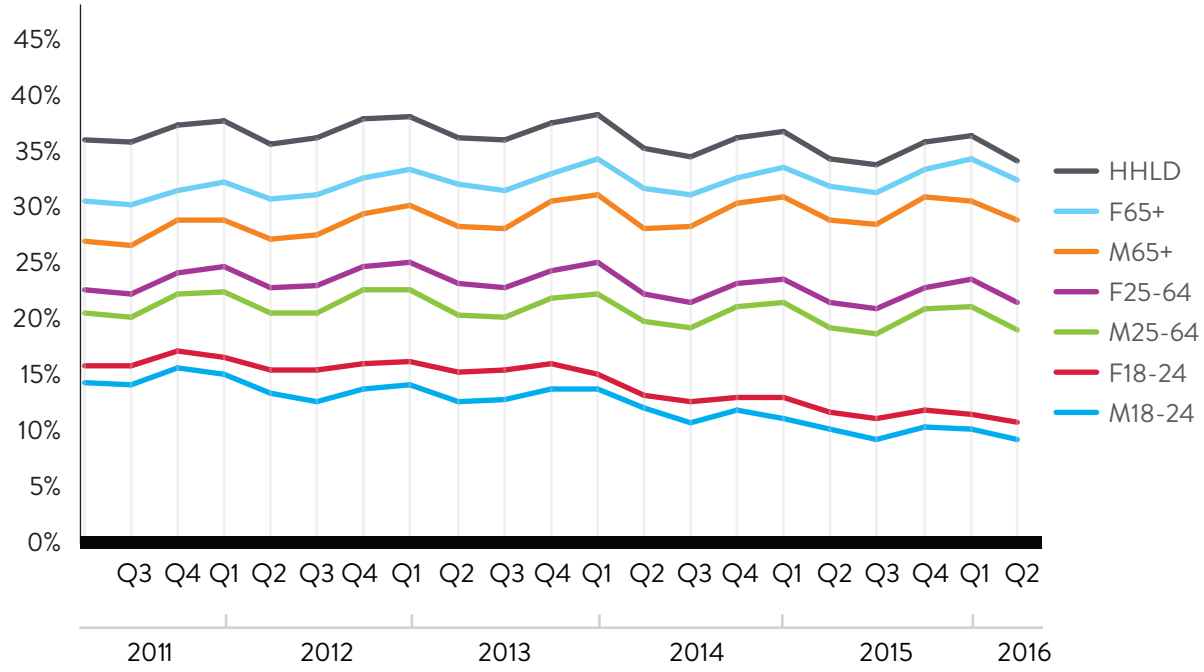


FIGURE 3: PERSONS USING LINEAR TV BY AGE AND GENDER



In another example (Fig. 4), preliminary analysis of time-shifted data for two specific networks—one broadcast network and one cable network—has allowed us to understand the rise of time-shifting activity over the years, and how much less seasonal that behavior has been in primetime for programs on the cable network, compared to programs on the broadcast network.

Those are just a few examples, but they illustrate the type of exploratory analysis that we performed to fully appreciate the scope, direction and overall quality of the data that we wanted to feed into our models.

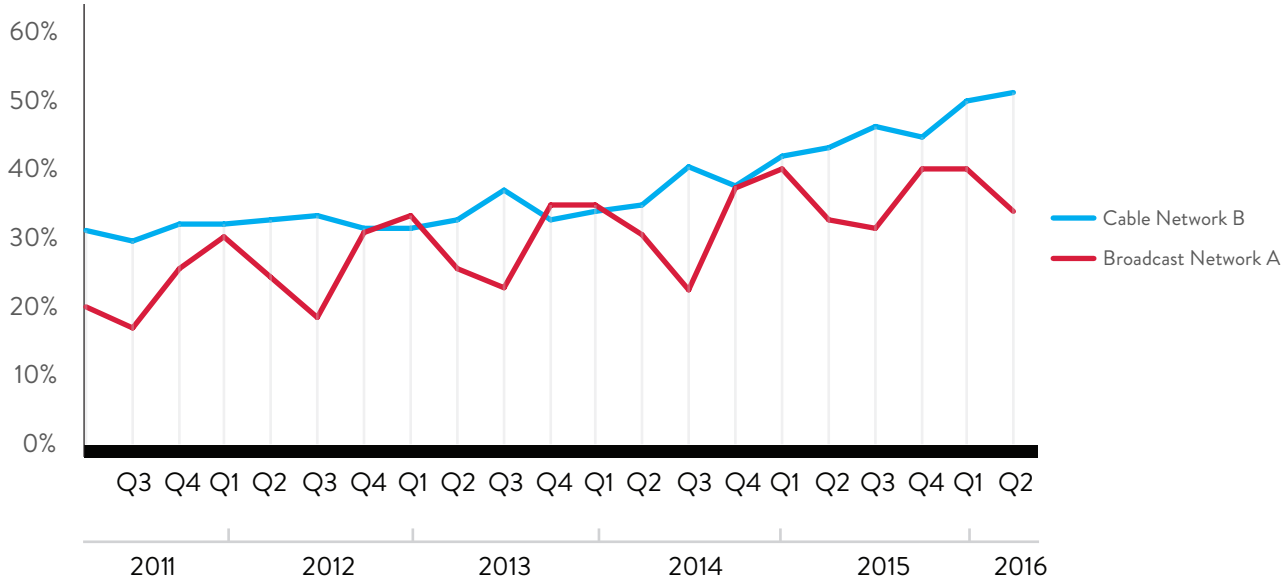
A DEEPER DIVE INTO OUR METHODOLOGY

In developing our projections, we tested many models and machine learning algorithms, including linear regression, penalized regression, multiple adaptive regression splines, random decision forests, support vector machines, neural networks and gradient boosting machine (GBM)[3]. While each method has its own advantages and disadvantages, in the end, the GBM method (specifically, the xgboost optimized library) proved to offer the best combination of accuracy and scalability for our project.

Gradient boosting is typically an ensemble (a model comprised of many smaller models) that utilizes many decision trees to produce a prediction. The illustration in Fig. 5 shows a simplified example of how an individual tree might work, and Fig. 6 shows how multiple trees might be aggregated in an ensemble to make a prediction.

We opted for xgboost, a recent variant of GBM, because it penalizes overly aggressive models—models that fit to the historical results too perfectly, a common mistake called "overfitting." Xgboost has taken the competitive prediction world by storm in recent years and frequently proves to be the most accurate and effective method in Kaggle[4] competitions. It's notably fast, scalable and robust.

FIGURE 4: RISE IN THE TIME-SHIFTED ACTIVITY FOR TWO SEPARATE NETWORKS



[3]*A discussion of the merits of each of these methods is beyond the scope of this paper. Interested readers will find a useful comprehensive resource in The Elements of Statistical Learning (by Hastie, Tibshirani, and Friedman).*

[4]*Kaggle is a crowdsourcing platform where data mining specialists post problems and compete to produce the best models. More information can be found at kaggle.com.*

FIGURE 5: A SIMPLE EXAMPLE OF A DECISION TREE

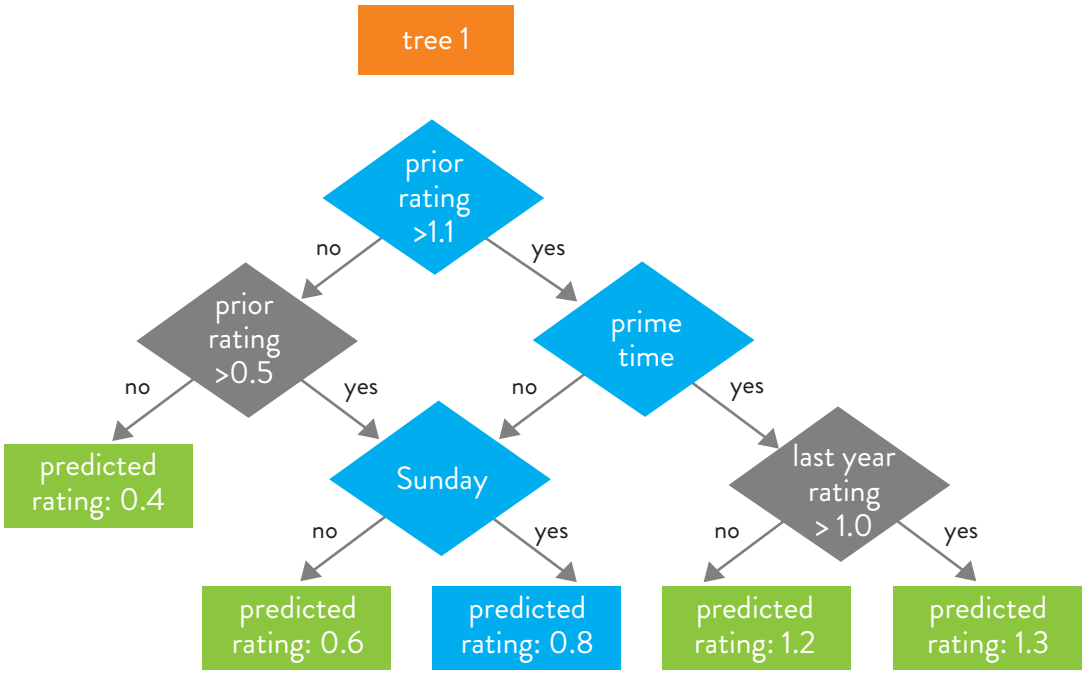
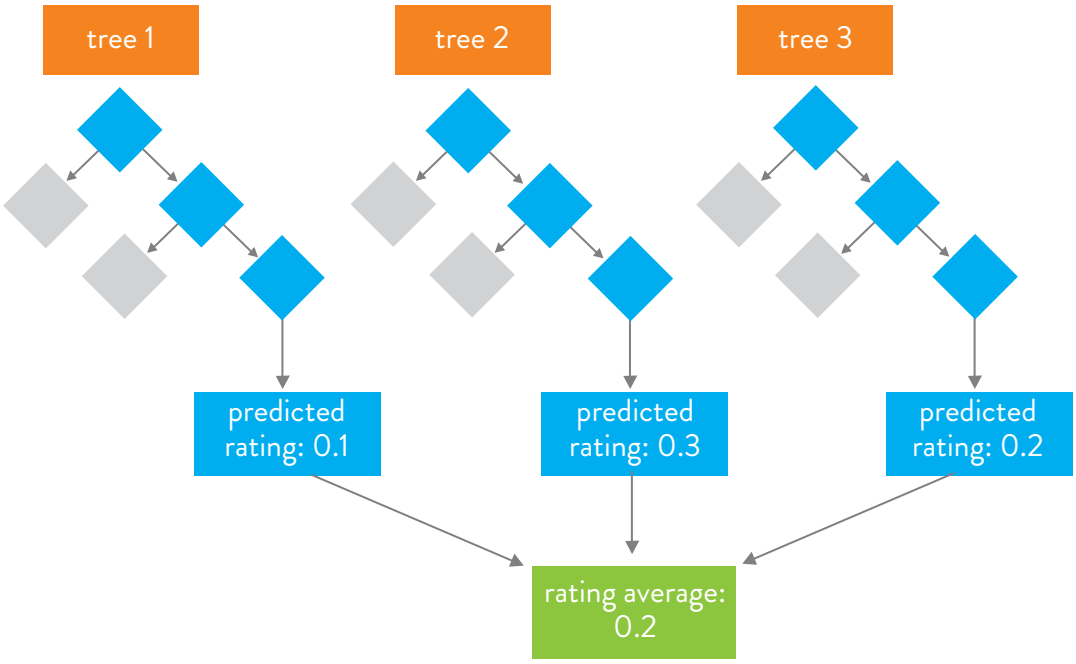


FIGURE 6: COMBINING MULTIPLE TREES INTO AN ENSEMBLE MODEL



SPLITTING THE DATA TO TRAIN A WELL-BALANCED MODEL

We restricted our data to only that which would be available when projections are typically made. Since the upfront occurs in May and June, it's technically possible for upfront projections to include some data from Q2, but for testing purposes, we decided to use only data through Q1 (and all relevant data from the preceding years, of course).

To be objective in assessing the accuracy of our projections, it was important to implement a fair and reliable process to develop our model and test our results along the way. Fig. 7 illustrates the iterative process we used to accomplish that goal.

Here are the main steps:

- Our algorithm randomly split the data into training and cross-validation testing sets. The model learned by making predictions based on the training set, testing those predictions on the cross-validation testing set, and repeating the process multiple times using different parameters. The final parameters were selected with consideration to the results of the cross-validation, helping limit the tendency to overfit the model to the training set.

- We also held out some data that was never used in the buildup process, but served as another layer to test the validity of our model and protect against overfitting. Holdout validation testing data provides an additional measure of quality control in the overall process. Models still tend to overfit even when using cross-validation. In order to choose the parameters most appropriate to apply to a new dataset, it is usually better to choose results that are slightly conservative, even for the testing dataset. The holdout validation testing set helped us achieve that balance.

- Once everything checked out and the final parameters were set, we retrained the model using the best parameters to leverage the most complete information available. We then ran it on a new dataset and compared its performance to client projections, focusing on key demographic groups.

FIGURE 7: AN ILLUSTRATION OF THE ITERATIVE PROCESS USED IN THE PROJECT

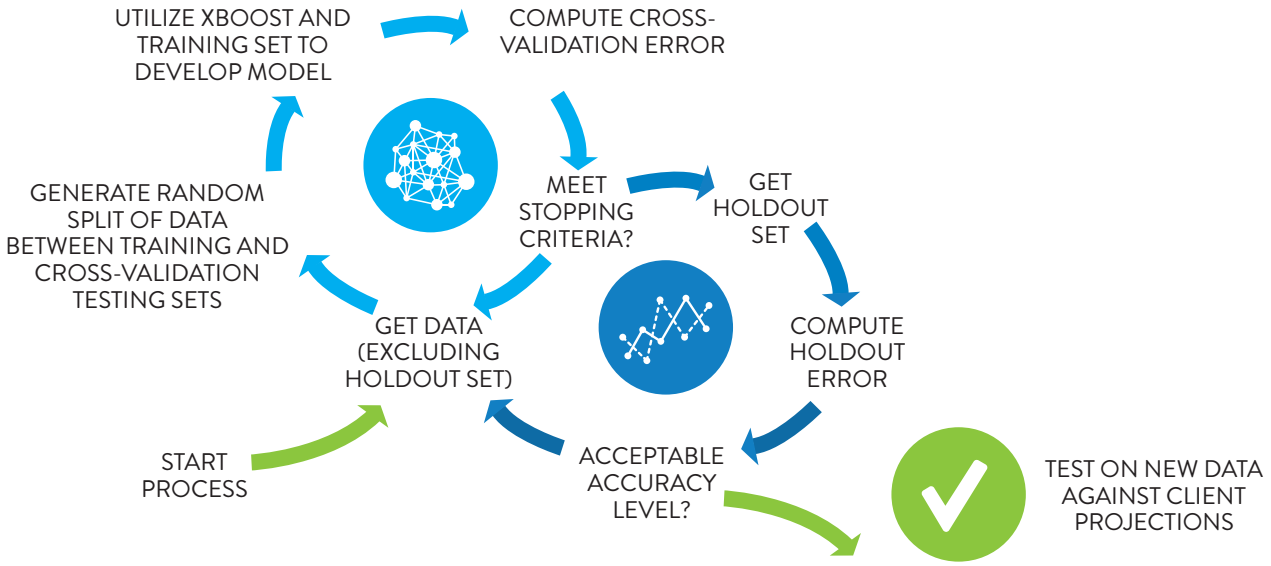
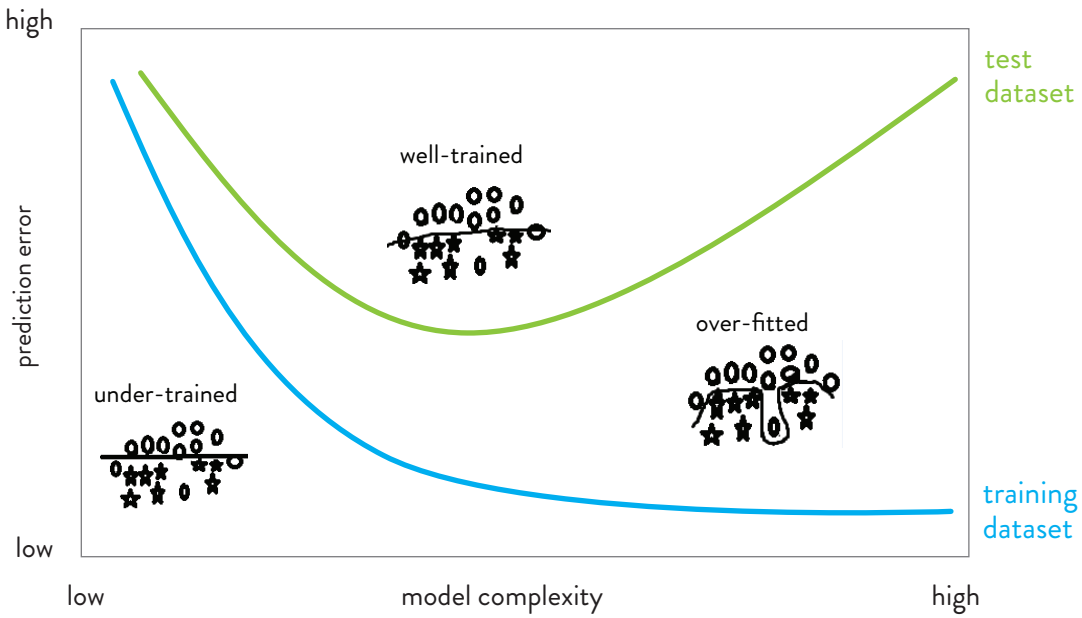


FIGURE 8: TRAINING ENOUGH WITHOUT OVERFITTING



We used cross-validation to build and evaluate our model. Cross-validation penalizes models that make predictions that fit too perfectly to past data, and thus are likely to reflect patterns that are too complex and unlikely to continue in the future. When training using cross-validation, we tried to find the point at which the model was able to capture important elements to make predictions, but ignored elements that were not powerful enough to offset the noise they created. The illustration in Fig. 8 can help visualize the point where a model starts to be too well trained for it to perform adequately on a new test dataset.

MEASURING THE PERFORMANCE OF OUR MODELS

As we evaluated our results, we focused on the following criteria:

- How close were our projections?

  We relied on a variant of WAPE (weighted mean absolute percentage error) to evaluate the accuracy of our models. WAPE is a statistical measure that helped us ensure that the way our model fit new data was reasonably consistent with how it fit historical data.

  We used WAPE to compare our model's accuracy to our client's model at two different levels. The first was at the channel level, which placed little emphasis on the ability to distinguish between programs, but was focused on getting the high level trends right—such as overall TV viewership for each channel. We also compared WAPE at the hour-block or program level. The hour-block level looked at the model's ability to distinguish between shows, as well as its ability to understand the high-level effects that influence all shows.

- How much information did the model explain?

  The metric of choice for this component was R-squared. R-squared is a statistical measure that represents the percentage of variation the model is able to explain. Unlike WAPE, R-squared did not evaluate if the high-level trends were captured appropriately. It was far more concerned with the ability to distinguish between programs, and was used to help establish the root of success or failure in our model at a more granular level.

- Was the model helpful?

  In addition to the hard evidence presented by WAPE and R-squared, we needed to consider practical implications of our process. For example, the model must be feasible for the client to implement. In addition, it should complement the client's existing framework. We also needed to identify where our projections could be trusted and when it might be more reasonable to lean on in-house knowledge. Finally, the accuracy of the model needs to be consistent enough to be trusted in the first place.

Our model was effective and produced several interesting findings. It held up close to expectations in terms of accuracy when evaluated using future testing dates. In addition, when we computed network performance using granular hour-block level data (we predicted 192 such observations for each network), our model's improvement over the client model was substantial for almost every network (see Fig. 9).

However, when we used aggregate network-level data (rather than hour-block level data) in our model, the results of our projections were far less clear. For some networks, we were closer, but for others, the client's model was more accurate in projecting the overall rating (see Fig. 10).

FIGURE 9: IMPROVEMENTS OVER CLIENT'S MODEL USING LOW-LEVEL OBSERVATIONS

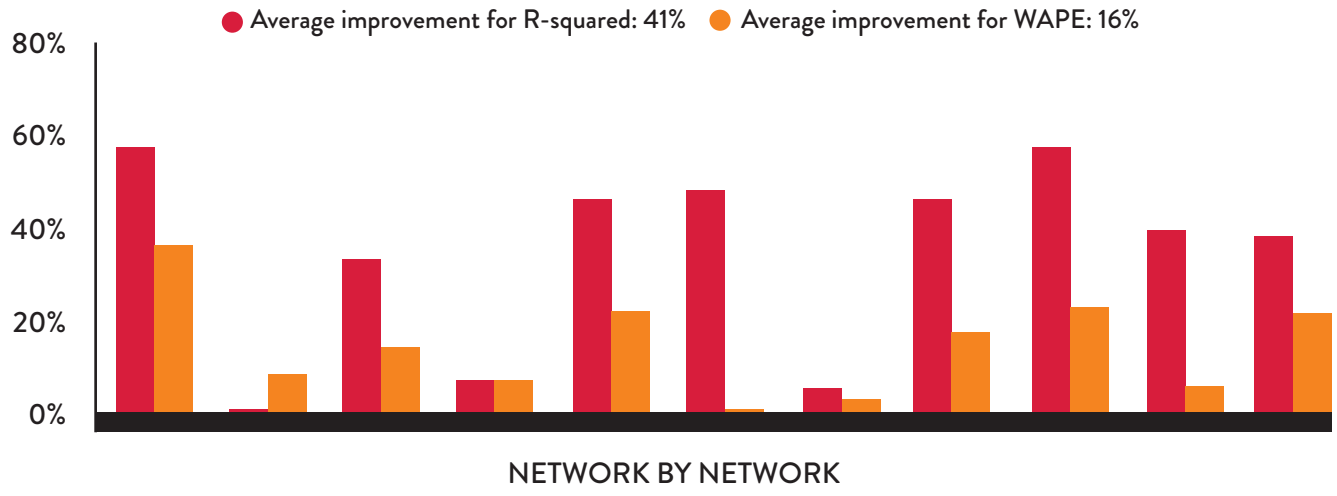
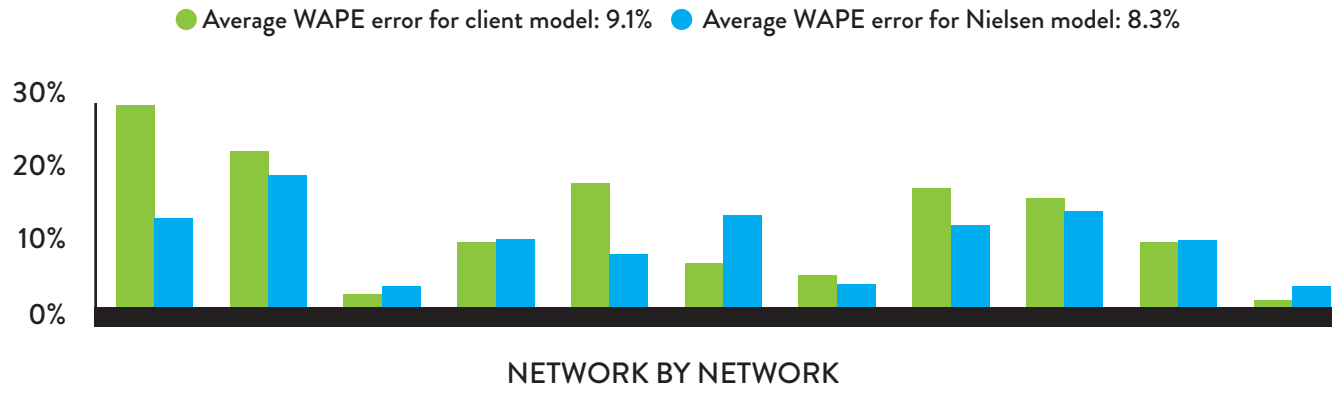


FIGURE 10: COMPARISON OF MODEL PERFORMANCE USING HIGH-LEVEL OBSERVATIONS



Why did the results look so different when rolled up to the network level? One possibility is that the client's model was able to capture unique in-house knowledge that could explain high-level effects that might have influenced all programs. It's also important to remember that a prediction at the network level relies on fewer prediction points, and might as a result be less reliable to begin with. We are probably very limited as to the conclusions that can be gleaned from the model at that level.

What is more interesting, however, is that when looking into the granular results for each network, we believe we see some indications as to how our model and the clients' projections might be combined to complement each other. First, we found that a model consisting of 90% our projection and 10% our client's projection outperformed each model individually in the two quarters that we tested. This was not isolated to just one case either: In fact, among the 11 regressions we ran for each of the channels, 10 suggested that both the client and Nielsen's models should contribute to a combined rating. This 90%/10% balance may not be the most robust estimate going-forward (as it should be validated over time), but it is certainly evidence that there is some unique knowledge contributed from both models.

Furthermore, there are some patterns that seem to emerge when we look at how each model complements the other from network to network. The network where a regression suggests the client's model contributes the most was rebranded and relaunched just five months after the upfront. This was somewhat expected given our prior assumption that the client's in-house knowledge should have more value when there are more significant changes taking place. To make this theory stronger, the network where a regression suggested that the client's model should have the second highest weight was rebranded and relaunched just before the upfront.

TOWARD A HYBRID MODEL

In the end, we were able to put together a robust model to predict future ratings, based on modern machine learning principles, and that model was particularly strong when the input data—and projected ratings data—was granular. However, for channels where we suspected in-house knowledge could play a key role, we found that the client's in-house model performed reasonably well. We believe that a hybrid model (one that can combine the raw power of our approach with custom insights) might be the best approach going forward.

There are additional benefits to combining forces. The time and energy required to generate thousands of projections are often beyond the resources of individual market research departments, especially for the lower-rated programs and time slots. An automated projection system can take care of the vast majority of projection estimates, and allow in-house experts to focus on the more important programs and factor in additional insights for those estimates. An in-house expert can also quickly evaluate the impact of unusual events and identify specific projections that are likely to go astray.

Of course, this doesn't mean that we shouldn't try and improve our predictive model: We might add more demographic characteristics to the model (e.g., income, location of residence, internet usage, etc.); Considering how much better our model performs with granular data than high-level data, we could take the analysis one step further and use respondent-level data; We might even add more competitive viewing data into the mix.

But the human element will always play a key role in the interpretation, so we might as well include that human element in the modeling process. The media landscape is changing fast, and those who are able to merge algorithms and intuition will be best positioned to anticipate the coming trends and capitalize on the opportunities.

nielsen