UNIVERSITY OF VIRGINIA
DATA SCIENCE INSTITUTE

Application of Machine Learning Methodologies to Multiyear Forecast of Video Subscribers

S&P Global
Market Intelligence

Jordan Baker, Andrew Pomykalski, Kaley Harahan, and Gianluca Guadagni

Abstract

The Market Intelligence division of S&P Global provides annual multiyear forecasts of United States subscribers for the video industry, which is comprised of cable, satellite, and telecommunication service providers. The current forecasts leave room for improvement as they are labor-intensive to generate and can be influenced by biases of subject-matter experts. The focus of this project is to explore the application of machine learning methods to these forecasts. We used demographic and subscriber data to assess the accuracy of different models, splitting it into training, validation, and testing sets. These models include support vector regression, artificial neural networks, and tree-based models, namely the traditional random forest and XGBoost. The training and validation sets were used to identify optimal parameters for each model. After identifying these parameters, we compared each model using the held-out testing set. These results indicated that the XGBoost model performed best, with a root mean squared error of 2,712. This initial exploration into machine learning shows promise and indicates it can be used as a tool to enhance S&P's current forecasting techniques. The industry is continuing to change as new technologies create disruption in the marketplace, which necessitates continued research into forecasting within this field.

Introduction

S&P Global's Market Intelligence division (SPGI) focuses on analyzing and interpreting global financial markets, such as video. This industry, comprised of cable, satellite, and telecommunication service providers, specifically relies upon the creation of forecasts that take into account current and historical trends and potential changes in the marketplace. These forecasts are consumed by a number of parties including financial institutions for investment evaluation, regulatory agencies for fair-play policies, and media conglomerates for strategic planning. With the knowledge of how the forecasts are consumed, there is a high importance placed on their accuracy.

The media and communications industry is transforming dramatically due to changes such as the increase in mobile technologies and over-the-top streaming services such as Netflix and Hulu. Traditional services such as cable are facing a decline while the demand for video through telecommunication providers is increasing [1]-[2]. With such change and uncertainty, the future of this industry is difficult to forecast.

Currently, forecasts produced by SPGI are generated by industry experts who take a holistic look at the factors that affect the video industry. While these experts have a vast amount of subject-matter knowledge, they undergo a labor-intensive and manual process to generate these forecasts. This provides the opportunity to explore the application of machine learning methodologies to assist SPGI's industry experts in forecast generation.

Goals and Objectives

- Explore the application of machine learning methodologies to forecasting video subscribers
- Understand the feature space of this problem and identify the features that are related to subscriber counts

Data

SPGI provided demographic data from 2007 to 2016, national-level subscriber data from 1995 to 2016, and county-level subscriber data from 2013 to 2016. The demographic data is broken down by county and service provider, and includes information on household incomes, gender and ethnicity, and general population features such as the number of households with only one vehicle. We offset the features from our response (video subscribers) by five years so that our model could be used to forecast five years into the future, given current demographic data. For example, demographics from 2010 will be paired with subscriber counts from 2015.

In order to make full use of the demographics, we used the national-level subscriber data, available back to 1995, to generate county-level subscribers for the missing years, 2007 to 2012. This required the calculation of the distribution of subscribers by county for 2013 to 2016 and for us to assume that these distributions are representative of previous years. We then distributed the national video subscriber totals across those counties.

2007 NATIONAL TOTAL = 98,030,643

County	Distribution	Subscribers
Albemarle	0.024%	24,071
Arlington	0.099%	97,945
Fairfax	0.302%	295,922

At the end of this process, we were left with five years of demographic data (2007 to 2011) paired with five years of subscriber data (2012 to 2016), resulting in 64,625 observations and 591 features.

Feature Space

With only demographic features available, we engineered a feature to capture shifts in the video marketplace. Market share was created by calculating the proportion of subscribers that a specific service provider has in a given county, for a given year.

To identify the features that impact subscriber counts we explored variable importance measures produced by a random forest. Variable importance is a measure of the decrease in node impurity as the trees are built, averaged over all trees.

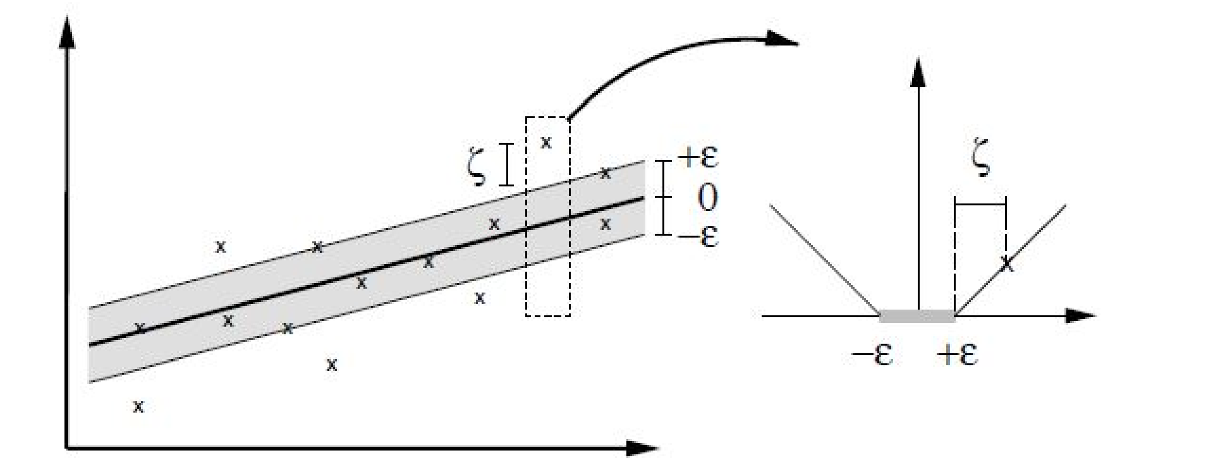
Feature	Importance
Market Share	0.393
Occupied Households (HH) with 1 Vehicle	0.074
Households	0.066
Non-Hispanic HOH, HHI $50k-$74k	0.066
1 Person Households	0.023
Female HOH, Non-Family HH, Ages 18+	0.023
Female HOH, Family HH, Ages 18+	0.014
Households with HHI $35k-$50k	0.013
2 Person Households	0.009
Population of Females Age 55-59	0.008

Machine Learning Techniques

We explored the effectiveness of four different machine learning techniques in forecasting video subscribers using county-level data. Each method was assessed using root mean squared error (RMSE) as the accuracy metric.

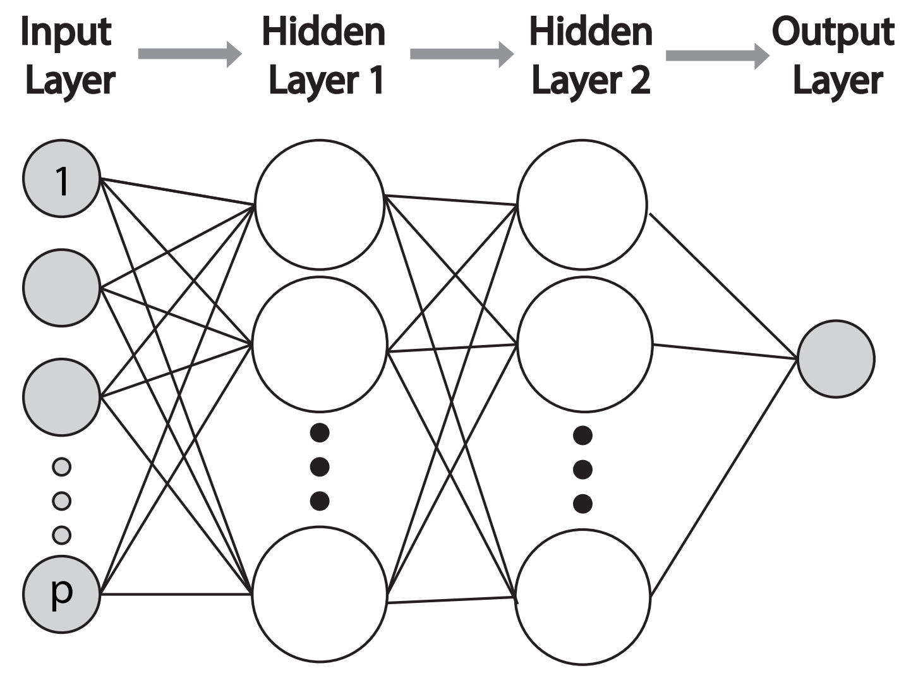
Support Vector Regression

Support vector regression (SVR) is a kernel method like support vector machines (SVM). The SVR algorithm separates the observations by constructing a function in the transformed, high-dimensional space [3]. There are two primary parameters that must be considered when working with SVRs: epsilon and cost. Epsilon, a nonzero value, represents the margins assigned around the function in which no penalty is assigned to incorrect predictions. Cost represents the penalty assigned to any incorrect predictions that fall outside the epsilon range and on the incorrect side of the hyperplane, as determined by class [4]. The kernel function minimizes the distance between observations and the epsilon region.



Artifical Neural Networks

A feedforward artificial neural network (ANN) is a biologically inspired learning method well equipped to learn patterns in complex, noisy data [5]. ANNs are comprised of layers of densely connected processing units that transform inputs, allowing for abstraction from the input vectors. Gradient descent backpropagation, a popular learning method, trains the network by updating the node weights iteratively to minimize MSE [5].



Tree-Based Models

A decision tree utilizes observations and features to build a series of if-then tests on features that result in a response value [5]. The tree is built starting at the root node, which is done by testing all features and then selecting the one that best separates the data into the given classes. The tree then iteratively continues to test features and build subtrees, expanding downward until all observations end in a terminal node.

A random forest is an ensemble modeling method that is made up of a large number of individual trees with slight modifications. These modifications include taking a bootstrapped sample of the data and using a random subset of the features each time a new tree is constructed [6]. Creating an ensemble of trees has been shown to provide a more accurate and stable prediction than a single decision tree, as well as fight overfitting [7].

A recent and popular extension of the random forest is the extreme gradient boosted random forest, or XGBoost. The fundamental concept behind XGBoost is the same as the traditional random forest, except XGBoost utilizes regularization to further reduce overfitting, improve accuracy, and decrease the time needed to construct trees [8].

Results

Model Comparisons

For each of these techniques we utilized a training/validation/testing approach to identify optimal parameters. A targeted grid search was used and multiple versions of each model were trained on the training set and compared their results when tested on the validation set to identify the optimal parameters. The parameters of the XGBoost model were identified using a 5-fold cross validated grid search. This type of grid search was possible due to the XGBoost model's ability to quickly create trees.

After identifying the optimal parameters for each model, a final RMSE was calculated on the held-out test set. This comparison metric represents the mean squared error between the forecasted value and the actual value, square rooted for interpretation purposes.

Machine Learning Model	RMSE
Support Vector Regression	16,365
Feedforward Neural Networks	9,262
Random Forest	3,515
XGBoost	2,712

XGBoost vs Legacy Model

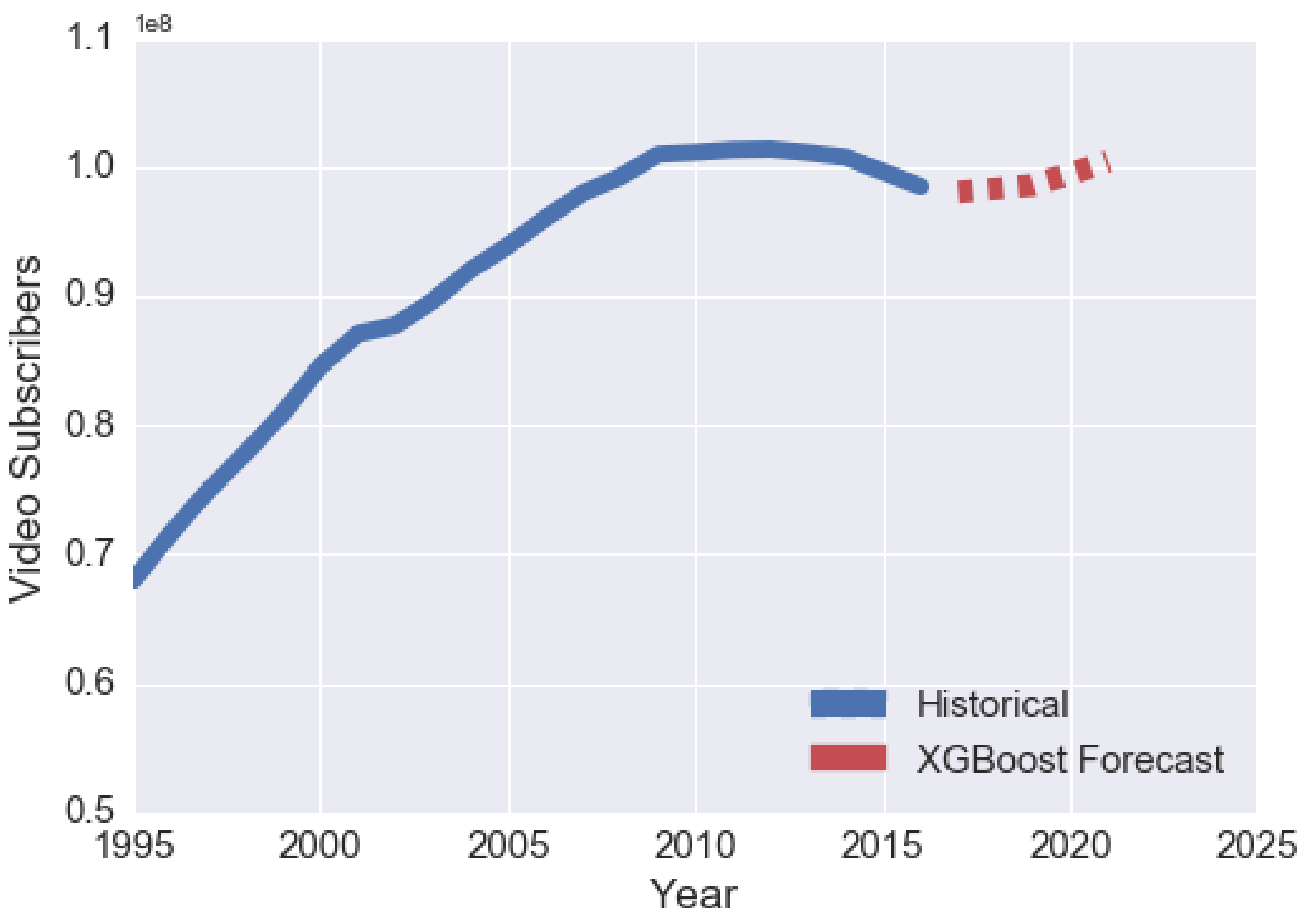
Comparing our optimal model with SPGI's existing method, we found that the optimal model produced forecasts for 2015 and 2016 that were more accurate than those produced by SPGI. SPGI's forecasts consistently predicted fewer subscribers than observed, while ours consistently predicted more.

Year	Actual	S&P		XGBoost	
		Forecast	Error	Forecast	Error
2015	99,687,914	97,900,000	*1,787,914*	100,174,770	*486,856*
2016	98,536,742	96,700,000	*1,836,742*	100,751,320	*2,214,578*
RMSE		1,812,492		1,603,338	

Forecast Results

To produce our final five-year forecast, we used the model selected through our training/validation/testing process, XGBoost. We trained the model on the full five years of subscribers and lagged predictors, then forecasted subscribers from 2017 to 2021.

Despite these recent changes in the overall trend, our model predicted an overall increase. Though these results may seem encouraging for machine learning methods, it is important to consider the extension of these methods as the final five-year forecast showed an increase in subscribers. This goes against the overall trend in recent years, which is showing a decline in video subscribers.



Conclusions and Future Research

The final exploration of machine learning methodologies showed that tree-based models performed best, with XGBoost outperforming the traditional random forest. Our model, combined with our exploration of the feature space, provides S&P Global's (SPGI) subject-matter experts with tools that can inform and enhance their current forecasting methods. This combination allows them to leverage machine learning methods as well as their advanced understanding of the industry.

There is substantial room for further research into forecasting the video industry. With such a dynamic industry, further incorporating its competitive nature would add value. As over-the-top services such as Netflix and Hulu grow, the entirety of their effect remains to be seen. Continuing to model this disruptive industry will require an adaptive approach.

References

[1] "Telecoms in 2017: A special report from The Economist Intelligence Unit." 2016. The Economist Intelligence Unit Limited.
[2] Lenoir, Tony. December 2015. "Connected Households Increasingly Opting out of Subscription TV." S&P Global Market Intelligence.
[3] Basak, Debasish, Pal, Srimanta, and Patranabis, Dipak C. 2017. "Support vector regression." Neural Information Processing-Letters and Reviews, 11(10), pp. 203-224.
[4] Smola, Alex J., and Schölkopf, Bernhard. 2004. "A tutorial on support vector regression." Statistics and computing, 14(3), pp. 199-222.
[5] Mitchell, Tom M. 1997. Machine Learning, Boston: The McGraw-Hill Companies. Regression by randomForest." R news, 2(3), pp. 18-22.
[6] Liaw, Andy, and Wiener, Matthew. 2002. "Classification and Regression by randomForest." R news, 2(3), pp. 18-22
[7] Srivastava, Tavish. June 2015. "Tuning the parameters of your Random Forest model." Analytics Vidhya.
[8] Jain, Aarshay. March 2016. "Complete Guide to Parameter Tuning in XGBoost." Analytics Vidhya.

Acknowledgements

We would like to thank Galen Warren, Ben Reneker, and Patrick Harrison of S&P Global Market Intelligence for their help and support throughout this project.