SMU Data Science Review

Volume 1 | Number 4 Article 3

2018

Improvements to Consumption Prediction: Machine Learning Methods and Novel Features

Ian Kinskey
Souther Methodist University, ikinskey@smu.edu

Glenn Oswald
Southern Methodist University, goswald@smu.edu

Charles McCann
Southern Methodist University, cmccann@smu.edu

Travis Finch
Southern Methodist University, tfinch@smu.edu

Anthony Tanaydin
Southern Methodist University, atanaydin@smu.edu

Improvements to Consumption Prediction: Machine Learning Methods and Novel Features

Ian Kinskey, Glenn Oswald, Charles McCann, Travis Finch, Anthony Tanaydin
Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
{ikinskey, goswald, cmccann, tfinch, atanaydin}@smu.edu

Abstract. Current models for predicting personal consumption expenditures (PCE) employ statistical techniques and rely upon traditional economic features. We compare vector autoregression and random forest regression models using traditional economic features as inputs to predict PCE. Additionally, we develop novel features derived from the earnings call transcripts of publicly traded U.S. companies using natural language processing (NLP) techniques. These new features reduce the mean square error (MSE) of the vector autoregression model by 7% and the random forest model by 23%. We find the random forest models outperformed the vector autoregression models, with a MSE reduction of 68%. We conclude the new features improve PCE predictions.

1 Introduction

U.S. Consumption (or personal consumption expenditures—PCE) represents 69% of gross domestic product (GDP). There is a long line of economic literature about the explanation and prediction of consumption. The majority of this literature focuses on the use of statistical regression methods using traditional economic measures as features. These economic measures, most of which are published on a monthly basis at a one month lag, attempt to capture data about investor and consumer confidence, the monetary and capital environments, the employment rate, and personal income. Few machine learning methods have been applied to the prediction of PCE. Additionally, while measures that attempt to quantify the current beliefs and future actions of investors and consumers are available, this sort of information is missing for the firms which supply the goods and services for consumption transactions.

In this paper, we compare a statistical forecasting technique, vector autoregression (VAR), which has been widely applied in the field of economics, to the random forest, a machine learning technique. We supply the same set of economic variables as inputs to each model—the three month treasury bill rate, the unemployment rate, personal income, the Index of Consumer Sentiment, and the New York Stock Exchange Composite Index. Additionally, we construct a new set of features which quantify and aggregate the sentiment of executives at firms selling consumer goods and services. These data and features are derived from transcripts of quarterly calls conducted by public companies to discuss financial results. We compare the performance of these features against the VAR and random forest models which only use the economic features.

A comparison of VAR models demonstrates a model which uses both the traditional economic features and the new executive sentiment features (the full model) yields a lower mean squared error (MSE) than the model which only uses the economic features (the reduced model). The reduced model MSE is $2,378 billion and the full model MSE is $2,202 billion, a 7% improvement. A similar comparison of random forest models yields the same results: the full model produces a lower MSE than the reduced model. The reduced random forest model MSE is $912 billion and the full random forest model MSE is $702 billion, a 23% improvement. A comparison of the full VAR model to the full random forest model shows the latter produces the lowest MSE. We thus conclude executive sentiment features improve PCE predictions.

The remainder of this paper is organized into six sections, references, and two appendices. Section 2 provides background and tutorial information on consumption, earnings of public companies, sentiment analysis, and provides a brief survey of the burgeoning field of natural language financial forecasting. Section 3 discusses data sources and collection methods. Section 4 outlines our analysis methods, and Section 5 examines the results. Section 6 provides an overview of ethical issues related to both this paper and NLP more generally. Section 7 offers conclusions and suggestions for future work. A listing of literature references and two appendices containing a sample earnings call transcript and a link to the program code conclude the paper.

2 Background and Tutorial

2.1 Personal Consumption Expenditure

Gross Domestic Product (GDP) is perhaps the most widely followed indicator of the health of the U.S. economy, and reports of GDP are closely follow by economists and investors. GDP can be computed and analyzed in terms of either production, income, or expenditures. [1] The production and income methods are irrelevant to the scope of this analysis. The expenditure view is defined as, "…the value of purchases made by final users—for example, the consumption of food, televisions, and medical services by households; the investments in machinery by companies; and the purchases of goods and services by the government and foreigners." [1] The expenditure method is more commonly measured, and is computed by summing government spending, private investment, net exports (exports less imports), and personal consumption expenditures. The values of each of these components and their corresponding proportions for the year 2017 are shown in Table 1.

Table 1. Components of GDP and their proportions for the year 2017.

Component of GDP	2017 Amount ($USD Trillions)	% of Total
Net Exports	-0.6	-3%
Government Spending	3.4	17%
Private investment	3.4	17%
Personal Consumption Expenditures	13.3	69%
Total GDP	**19.5**	**100%**

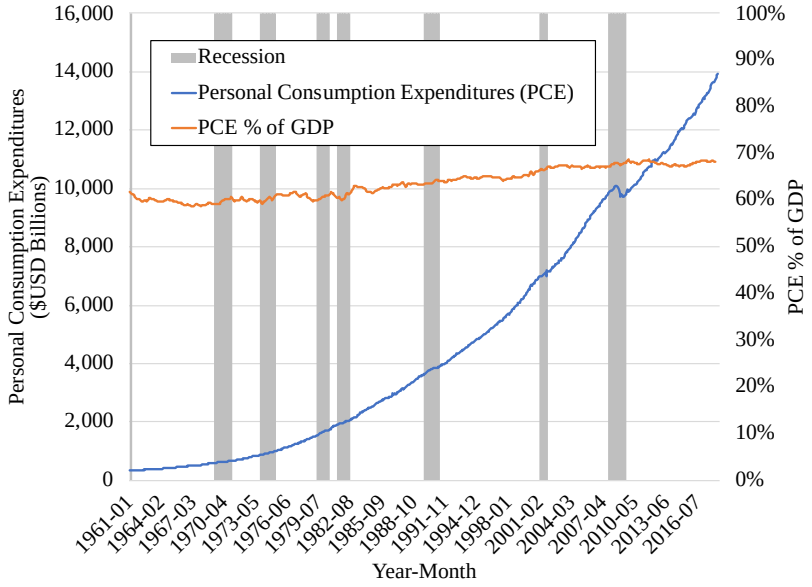


Fig. 1. Personal consumption expenditures by month (1961 to 2018). PCE accounted for nearly 70% of GDP in 2017.

The largest component of GDP is personal consumption expenditures (PCE), which accounted for 69% of GDP [2] in 2017 as shown in Figure 1. The Bureau of Economic Analysis (BEA) defines PCE as, "…the value of the goods and services purchased by, or on the behalf of, U.S. residents." At the national level, BEA publishes annual, quarterly, and monthly reports containing estimates of PCE. [3] The goods included in the PCE calculation comprises both durable and non-durable goods. Durable goods include motor vehicles, furniture and household equipment, and recreational goods and vehicles. [2] Non-durable goods include food and beverages, clothing and footwear, and fuel and energy. [2] The services included in the PCE calculation include housing and utilities, health care, transportation services, recreation services, food services and accommodations, financial services and insurance, and other miscellaneous services.

2.2 Earnings Calls

A business entity is generally referred to as a public company if its shares are listed for sale on a public stock exchange and it discloses financial results regularly to the public. The Securities and Exchange Commission (SEC) Division of Corporate Finance requires that public companies release periodic financial results. [4] Beyond the SEC requirements, companies are motivated to report financial results to maintain communication between the shareholders, boards of directors, management, other stakeholders, and potential investors of those companies.

Public companies in the U.S. typically release financial results every three months (quarterly). Disclosure and reporting of financial results typically adheres to the following sequence: (1) the public is notified in advance of the upcoming earnings release, (2) a summary of financial results including earnings are published ("earnings release") via press release and the company website, (3) the company conducts an earnings conference call ("earnings call") to discuss financial results, and (4) the company files formal SEC documents. An excerpt from an earnings call transcript is provided in the first appendix section of this paper.

An earnings release, prepared by a company's management, is a preview of the financial results that will be reported in the formal regulatory filing required by the SEC. At a minimum, revenue, operating margins, net income, and earnings-per-share are typically presented. An earnings call is a live review of the earnings release and is management's opportunity to add insight to the raw numbers. Management can address other qualitative or quantitative information that would explain events in the previous period. Management participation in a conference call varies by company but usually involves one or more senior executives, and often includes both a company's chief executive officer (CEO) and chief financial officer (CFO). Industry research indicates that 97% of public companies now conduct earnings calls. [5]

An earnings call has two distinct sections. The first section is a presentation of the results of operations and financial performance of the company. This section can either be broadcasted live or pre-recorded but is usually a scripted recitation of the financial results published in the earnings release. The second section is an extemporaneous question and answer section where management responds to questions from industry analysts about those results. This is a live exchange where management responds to questions in real time, communicating both facts and opinions about the company.

Management often gives guidance on their expectations for future periods. Shareholders and industry analysts place significant importance on these forward-looking statements. The National Investor Relations Institute identified some important statistics about forward looking statements in their *Earnings Process Practices Research Report* published in 2016. This report concluded that 94% of companies provide guidance for future periods. [6] Guidance often communicates both financial and non-financial information. Management provides opinions on expectations for their own company, the industry in which they operate, and the general economy.

Earnings calls typically occur 30 to 60 days after the end of the previous period and thus occur well into the subsequent period of operation. Therefore, management possesses substantial information about the potential for the results of the current quarter at the time of the earnings call.

2.3 Sentiment Analysis

Sentiment analysis, sometimes called opinion mining, is the set of natural language processing techniques used for extracting and quantifying the emotional, affective, tonal, and sentiment information in a text. [7] Sentiment analysis can be performed on a text consisting of a single sentence, a full document, or a set of documents. While the subfield of sentiment analysis is rooted in linguistics research dating back decades, the application of data science techniques did not begin to take hold until around the year 2000. [7] At present, sentiment analysis is one of the most popular areas of data science research in both academia and industry. Although there are many kinds of sentiment analysis, we focus here on sentiment lexicon based techniques. See Liu's *Sentiment Analysis and Opinion Mining* [7] for a thorough exploration of the topic, including other sentiment analysis methods and techniques. A sentiment lexicon is a list of words and phrases highly associated with some expression of sentiment. These lexicons are manually developed by researchers. Texts are then searched for the occurrence of these words and phrases and their frequencies are tabulated to compute measurements of sentiment.

The Loughran-McDonald (LM) sentiment lexicon was developed from an analysis of SEC filings. [8] Loughran and McDonald showed that lexicons developed for other domains performed very poorly on texts from the finance domain, with some popular lexicons misclassifying 74% of positive finance words as negative. [8] Loughran and McDonald mined over 50,000 SEC filings spanning 1994 to 2008 to create their lexicon, which includes a list of 85,131 words resolving to 62,374 lemmas (i.e., uninflected base forms of words). [8] Of these lemmas, 3,141 are categorized as having sentiment value in one of seven categories: negative, positive, litigious, uncertain, modal weak, modal moderate, and modal strong. Table 2 provides definitions, counts, and examples for each sentiment category.

Table 2. Sentiment categories from the Loughran-McDonald lexicon [8]

Sentiment Category	Definition	Example Words
Negative	Words denoting a negative tone, sentiment, or outlook.	Loss, Against, Claims, Termination, Impairment, Adverse, Failure, Default
Positive	Words denoting a positive tone, sentiment, or outlook.	Effective, Benefit, Able, Gains, Greater, Good, Best, Beneficial
Litigious	Words reflecting a propensity for legal contest or litigiousness.	Shall, Herein, Amend, Thereof, Contracts, Law, Claims, Legal, Laws, Hereof
Uncertain	Words denoting uncertainty, emphasizing the general notion of imprecision rather than exclusively focusing on risk.	Approximately, Risk, Believe, Believes Assumptions, Intangible, Anticipated
Modal Weak	Words expressing weak levels of confidence.	Could, Possible, Might, Depending, Appears, Nearly, Sometimes, Almost
Modal Moderate	Words expressing moderate levels of confidence.	Would, Generally, Can, Should, Likely, Probable, Often, Regularly, Frequently, Usually
Modal Strong	Words expressing strong levels of confidence.	Will, Must, Best, Highest, Never, Lowest, Always, Clearly, Strongly, Undisputed

2.4 Natural Language Financial Forecasting

Natural language financial forecasting (NLFF) combines the disciplines of linguistics, natural language processing, machine learning, statistics, and behavioral economics. [9] The literature of this developing field begins in earnest with Wuthrich (1998), wherein daily directional changes of five major stock indices are predicted using text from fifteen different financial news sources as inputs. [10] Since then, the field of NLFF has seen significant growth in research publications, as shown in Figure 2.

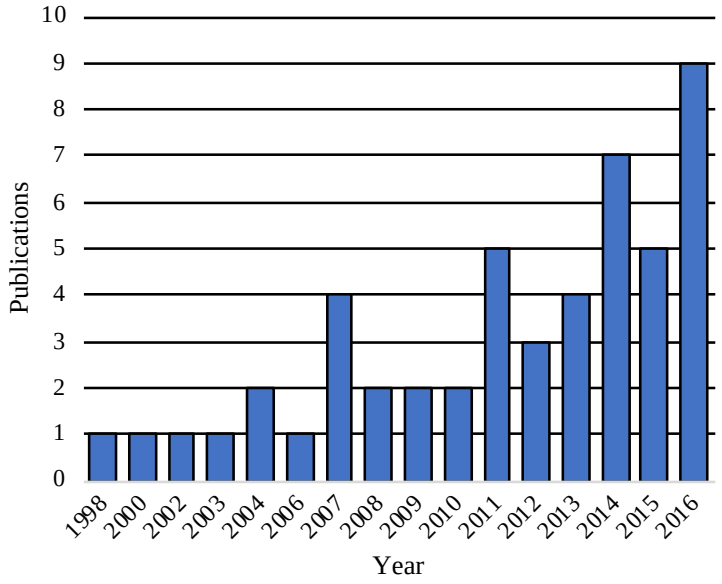


Fig. 2. The frequency of research papers published in the NLFF field by year of publication.

Data utilized for NLFF fall into two categories: text (non-numeric) data and market/economic (numeric) data. Text data are used as features and are generally mined from five different source types: (1) corporate filings and disclosures such as SEC filings and published annual reports, (2) financial reports compiled by third party financial researchers (e.g., investment banks, credit rating agencies, and other financial market participants), (3) financial news and periodicals such as Bloomberg, The Wallstreet Journal, Financial Times, Bloomberg, Thomson Reuters, Yahoo! Finance News, Dow Jones, Forbes, et cetera, (4) social media which is comprised entirely of data collected from the microblogging web application Twitter, and (5) internet message boards such as Yahoo! Finance Message Boards, Raging Bull, and The Motley Fool Message Boards. [9][11] Figure 3 depicts the frequency and proportions of text source used by publications. Notably, none of the surveyed literature utilize earnings call transcripts as part of their analysis. The number of documents collected for each paper ranges from hundreds to tens of millions and span time ranges from weeks to decades. These text features are sometimes combined with other more traditional data

sources such as the historical prices of financial instruments and/or the financial statement data related to those financial instruments.

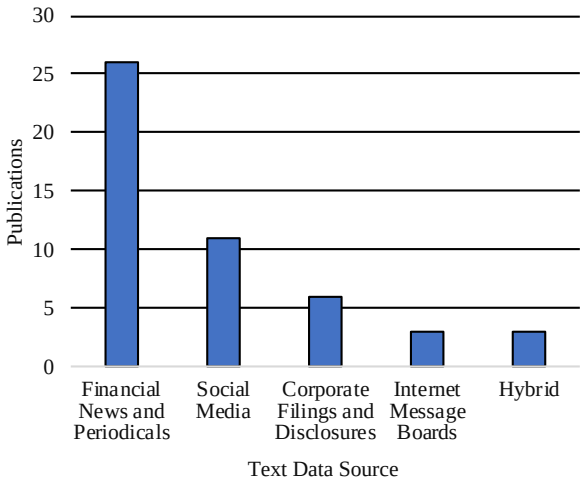


Fig. 3. The frequency and proportions of text data sources used in the NLFF literature. Hybrid sources entail the combination of two or more of the other source types. For example, a hybrid source type would be the use of both "Financial News & Periodicals" and "Social Media".

The NLFF literature uses a variety of algorithms to make financial predictions. Figure 4 displays the distribution of prediction algorithms used in the NLFF literature over time. These models generally take as input the results of NLP models and, in some cases, other data such as financial fundamentals, economic data, and market data. The support vector machine family of machine learning algorithms is the most consistently and commonly used in NLFF due to their ability to easily handle a large number of features, computational efficiency, and scalability. Linear regression is also quite popular, owing to the method's ease of use and interpretability. Finally, neural networks of various types are seeing significant usage growth in recent years, matching their increased usage of in the wider machine learning and artificial intelligence (AI) domains.

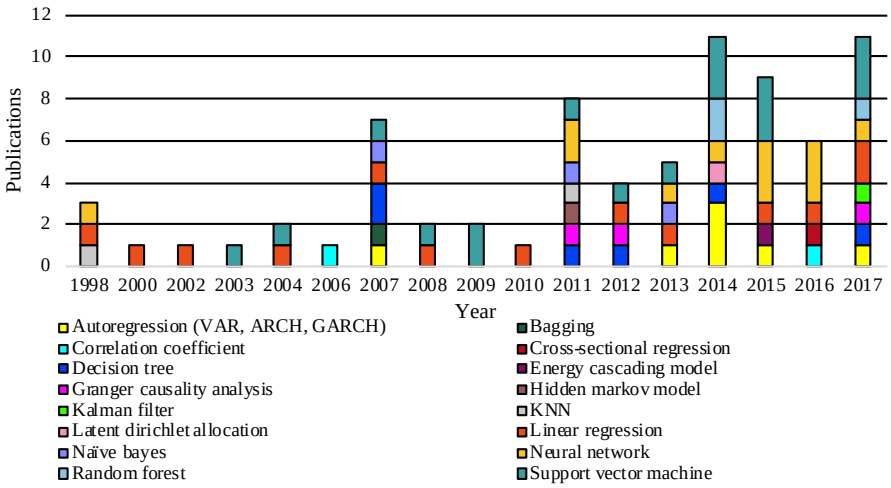


Fig 4. Prediction algorithms used in the NLFF literature. Many publications use more than one algorithm and are counted multiple times.

Within the NLFF literature, the metrics used to evaluate models vary widely depending on the types of predictions. Metrics for success of prediction fall into three categories within the literature: directional accuracy (classification accuracy), closeness of the predicted value to the actual observed value (regression error), and trading simulation results. [11] For directional accuracy evaluation metrics a prediction is made about the category of future results such as whether a stock price will go up or down in a future period but do not predict the magnitude of change. [11] The objective of studies using this type of evaluation metric is to maximize accuracy. The accuracy of those predictions is then computed as the proportion of predictions which were correct. [11] For regression error evaluation metrics, the difference between predicted and actual values is computed with an objective of minimizing these errors. Common metrics of this type includes mean square error (MSE), root mean square error (RMSE), and mean absolute percentage error (MAPE). [11][12] The third class of evaluation metrics, trading simulation results, are derived from hypothetical trading transactions suggested by the predictive model. The success of these simulations is evaluated based on the resulting hypothetical profits and metrics include average percentage gain per transaction (APGT), the accumulated return of the transactions, and traditional trading metrics which compare the risk and reward of the transactions such as the Sharpe ratio. [11]

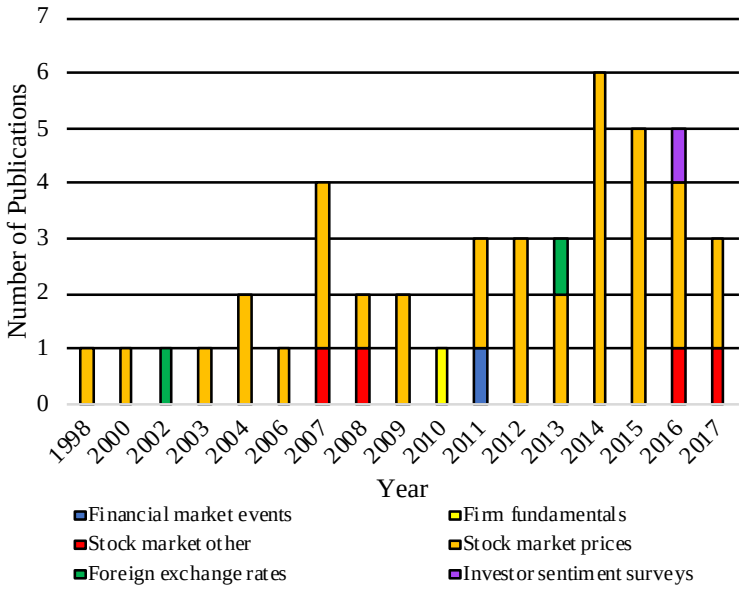


Fig. 5. Count of NLFF publications by year, categorized according the type of predicted/response variable studied.

Figure 5 depicts the types of response variables which are predicted or whose relationship with NLFF text data is examined. These predicted variables breakdown into six categories: (1) financial market events such as mergers, acquisitions, stock splits, et cetera, (2) other stock market variables corresponding to either individual firm securities or market indices such as trading volume and volatility, (3) foreign exchange rates (e.g., U.S. Dollars to Euros), (4) firm fundamentals (gross revenue, net income, et cetera), (5) stock market prices of either individual firms or stock indices, and (6) investor sentiment surveys. As Figure 5 shows, the majority of NLFF literature focuses on the prediction of stock market prices.

3 Data

3.1 Macroeconomic Measures

We obtained data for various economic variables as shown in the Table 3. All data are of the numeric type float and are continuous variables. Additionally, all data span the time frame of November 2010 to June 2018 and are of a monthly frequency. The 3-Month Treasury Bill, Secondary Market Rate data represent the interest rate of the U.S. federal government's short term debt obligation. The secondary market interest rate of this security is an indicator of the monetary policy climate, inflation, and investor

confidence. The NYSE Composite Index is a composite index of the prices of all the companies traded on the New York Stock Exchange. The unemployment rate is the percentage of the individuals in the U.S. labor force over the age of 16 who are unemployed but actively seeking employment. This measure is commonly referred to as the "U-3" unemployment rate. The Index of Consumer Sentiment is derived from consumer surveys conducted by the University of Michigan's Institute for Social Research. The surveys seek to measure how consumers feel about their immediate financial future and the state of economy. Personal income is a measure of the aggregate personal income of individuals in the U.S. Personal income is used as a proxy for the financial health of individuals in the United States and their capacity for spending.

Table 3. Descriptions of macroeconomic data used in this analysis.

Variable	Source	Units
3-Month Treasury Bill Secondary Market Rate	St. Louis Federal Reserve	Percent
NYSE Composite Index	Yahoo! Finance	Index
Unemployment Rate	U.S. Bureau of Labor Statistics	Percent
Index of Consumer Sentiment	U. of M. ISR	Index
Personal Income	U.S. Bureau of Economic Analysis	Billions of U.S. Dollars
Personal Consumption Expenditures	U.S. Bureau of Economic Analysis	Billions of U.S. Dollars

3.2 Earnings Call Transcripts

An earnings call transcript is a transcription of a company's quarterly earnings call. We scrape approximately 150,000 transcripts from the website SeekingAlpha.com. We use the scraping packages ScraPy and BeautifulSoup via the Python programming language. We then construct a sample frame of 610 companies, limiting eligibility to those companies trading on major U.S. exchanges (New York Stock Exchange, NYSE American Exchange, or NASDAQ) and who are members of either the "Consumer Defensive" or "Consumer Cyclical" sectors. These two sectors are chosen as their customers are almost exclusively consumers and therefore the majority of the revenue of each company directly contributes to PCE. We partition the sample frame into five strata according to the most recently reported annual revenues of each company. Finally, we draw a random sample of 50 companies from this sampling frame, with the sample size allocated to each stratum via the Neyman allocation method.

4 Methods

4.1 Sentiment Analysis

Earnings calls follow a quarterly cadence, so we first assign each transcript to a period of three consecutive months. Each company's transcripts are assigned to the calendar month in which they first became available and the two subsequent months.

Next, we compute sentiment and other natural language processing measures using both the Loughran-McDonald sentiment lexicon and a parsing and computation program developed by those same researchers. This program counts the occurrences of words which have been identified as corresponding to the LM lexicon sentiment categories: positive, negative, uncertain, litigious, modal weak, modal moderate, and modal strong. The parsing program computes additional statistics from each transcript document, all of which are shown in Table 4. Finally, we aggregate each of these measures by quarter, creating two sets of variables according to method of aggregation. The first set are the variables shown in Table 4, aggregated via simple arithmetic mean. The second set are the same variables, but aggregated via a weighted arithmetic mean with weight corresponding to the most recently reported annual revenues of each company associated with each transcript. For example, Walmart, with 2017 revenues of $500,343,000,000 would have an influence 1,649 times that of U.S. Auto Parts whose 2017 revenues were $303,366,000.

Table 4 Variables computed from earnings call transcripts.

Variable Name	Description
numWords	Number of words in transcript
percPos	% of words in transcript which are in LM's positive word list
percNeg	% of words in transcript which are in LM's negative word list
percUncert	% of words in transcript which are in LM's uncertain word list
percLitig	% of words in transcript which are in LM's litigious word list
percModWk	% of words in transcript which are in LM's modal weak word list
percModMd	% of words in transcript which are in LM's modal moderate list
percModSt	% of words in transcript which are in LM's modal strong word list
percConstr	% of words in transcript on LM's list indicating financial constraint
numAlphaNum	Number of words in transcript that are alphabetic and/or numeric
numDig	Number of words that are single digit numbers
numNum	Number of words of any length that are numeric
avgSyll	Average number of syllables per word in transcript
avgLen	Average length of words used in transcript
vocab	Number of unique words used in transcript

4.2 Vector Autoregression Models

After reviewing the literature covering prediction of PCE, we choose to adapt a vector autoregression (VAR) model developed by E. Philip Howrey. [16] This model is selected for its success in predicting PCE, its relatively simple methodology, and the ease by which updated data are obtained. This model uses the economic variables described in section 3.1.

VAR models are a specific type of multivariate time series capable of predicting future values of multiple variables. VAR models include lagged values of the response variable as a regressor (endogenous variable). Other variables (exogenous variables) may be included as regressors in the model, which are also lagged to previous values. In a VAR model each variable is a linear function of the lagged values of itself as well as the lagged values of the other variables in the model. The order of the model determines how many series of lagged value terms are included. VAR models follow the general form show in Formula 1.

$$\underset{k\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}} = \underset{k\times 1}{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}} + \underset{k\times k}{\begin{bmatrix} w_{1,1} & \cdots \\ w_{2,1} & \vdots \\ \vdots & \\ w_{k,1} & w_{k,k} \end{bmatrix}} \underset{k\times 1}{\begin{bmatrix} y_{1,(t-1)} \\ y_{2,(t-1)} \\ \vdots \\ y_{k,(t-1)} \end{bmatrix}} + \cdots \underset{k\times k}{\begin{bmatrix} w'_{1,1} & \cdots \\ w'_{2,1} & \vdots \\ \vdots & \\ w'_{k,1} & w'_{k,k} \end{bmatrix}} \underset{k\times 1}{\begin{bmatrix} y_{1,(t-p)} \\ y_{2,(t-p)} \\ \vdots \\ y_{k,(t-p)} \end{bmatrix}} + \underset{k\times 1}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{bmatrix}} \quad \textbf{(1)}$$

The vector equation above represents a *p*-lag vector autoregressive model *VAR(p)*. The *p*-lag is the number of previous periods' (lags) values which are used as inputs. On the left of Formula 1 is a $(k\times 1)$ vector of present time period response variables y_1 through y_k. Next is a $(k\times 1)$ vector of constant terms as the intercepts. Each variable in the model is represented as a $(k\times 1)$ vector of lagged terms $y_{1,(t-1)}$ through $y_{k,(t-1)}$ preceded by a $(k\times k)$ matrix of coefficients. Additional variable vectors and coefficient matrices are added to represent the model order up to time *t-p*. Finally a $(k\times 1)$ vector of irreducible errors (ϵ) that have mean zero and are uncorrelated and independent.

Howrey modeled a period extending from 1978 to 2001. We update the time frame to extend from January 2011 to June 2018, and change the period from quarterly to monthly. We use Baysean information criterion (BIC) to select a lag order of 1 for our updated VAR. These updates yield the reduced model.

As one means of evaluating whether the executive sentiment data improves the ability to predict PCE, we create a full VAR model, which, in addition to using the economic variables shown in Table 3, also uses our variables of executive sentiment and other NLP measures of earnings call transcripts as shown in Table 4. Similarly to the reduced VAR model, the lag order for the full model is selected using BIC.

4.3 Random Forest Models

The random forest algorithm is a supervised machine learning algorithm capable of performing regression and classification tasks. This algorithm creates a collection of decision tree models trained on data sets bootstrapped from a complete, original training data set, a concept known as bootstrap aggregation (i.e., "bagging").

Predictions are aggregated from the predictions of the set of models via a simple arithmetic mean. The algorithm proceeds along the steps shown in Figure 6.

```
1  Input the number of decision trees hyperparameter, t, to train
2  for 1 to t
3  |    Randomly select with replacement a sample from the training set of
   |    equal size to the original training set (the bootstrap sample)
4  |    Train the decision tree model.
5  |    while the magnitude of the decision tree model MSE continues to
   |    decrease, do
6  |    |    Randomly select k of the P features of the data, where
   |    |    k < P
7  |    |    Determine which of the k features produces the best
   |    |    split using the selected feature
8  |    Split on the best feature
9  |    end
10 end
```

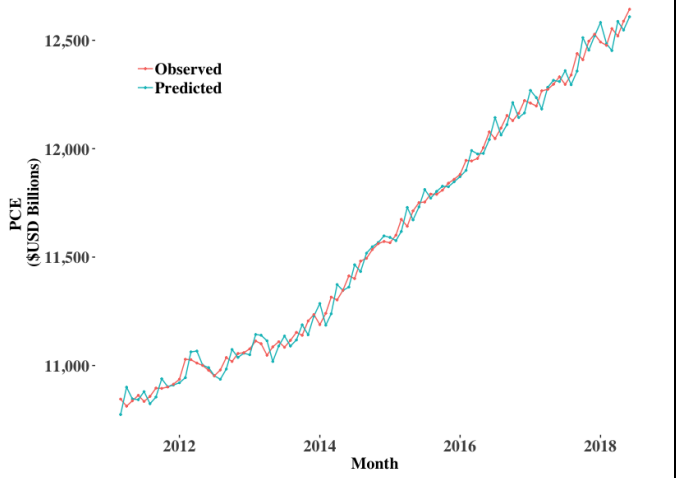
Fig. 6. The random forest algorithm. [17]

We train two random forest models. The first model uses only economic variables as features (the reduced model), and the second model uses those same economic variables as well as our variables of executive sentiment and other NLP measures extracted from earnings call transcripts (the full model).

5 Results

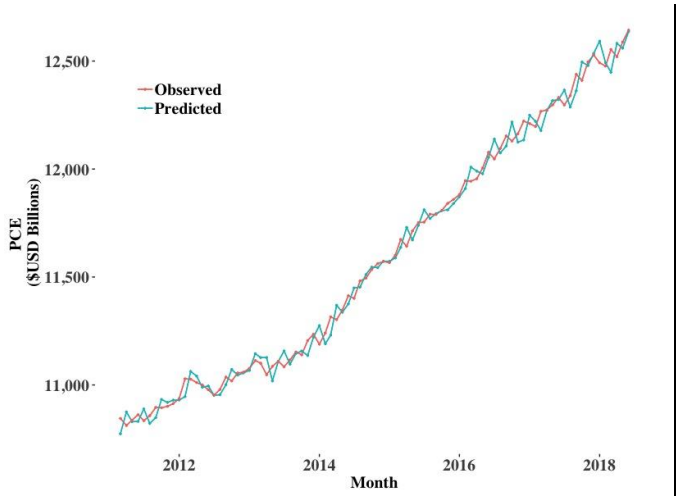
While we display multiple metrics for each model, we use mean squared error (MSE) as our primary evaluation metric.

5.1 Vector Autoregression Comparison



MSE	$2,378B
RMSE	$49B
Adj. R^2	0.2407
p-value	6.7e-05

Fig. 7. Plot of monthly observations and predictions by the reduced VAR model of PCE, as well as metrics of that model.

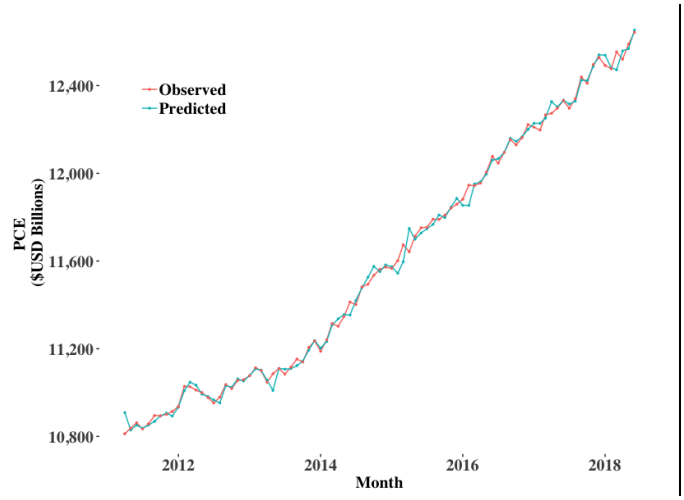


MSE	$2,202B
RMSE	$47B
Adj. R^2	0.2244
p-value	2.2e-03

Fig. 8. Plot of monthly observations and predictions by the full VAR model of PCE, as well as metrics of that model.

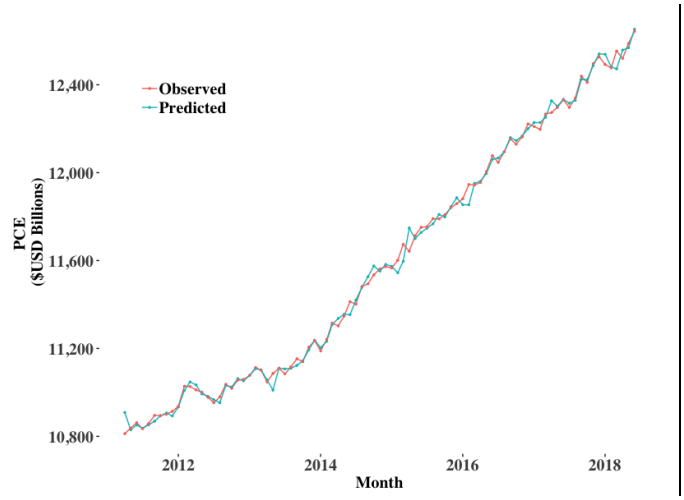
To compare the two VAR models we perform a likelihood ratio test. The likelihood ratio test is used to compare nested models and identifies the model which is most likely given the parameter estimates of each model. This test yields a p-value less than 0.001, providing strong evidence that the full VAR is superior to the reduced VAR.

5.2 Random Forest Comparison



MSE	\$912B
RMSE	\$30B
Adj. R^2	0.4271

Fig. 9. Plot of monthly observations and predictions by the reduced random forest model of PCE, as well as metrics of that model.



MSE	\$702B
RMSE	\$27B
Adj. R^2	0.4722

Fig. 10. Plot of monthly observations and predictions by the full random forest model of PCE, as well as metrics of that model.

To compare the models we randomly sample observations from our data set and train full and reduced random forest models. This exercise is repeated 50 times, and we calculate the MSE for each model. Figure 11 shows a boxplot comparing the distributions of mean square error metrics computed for each iteration of reduced and full models.

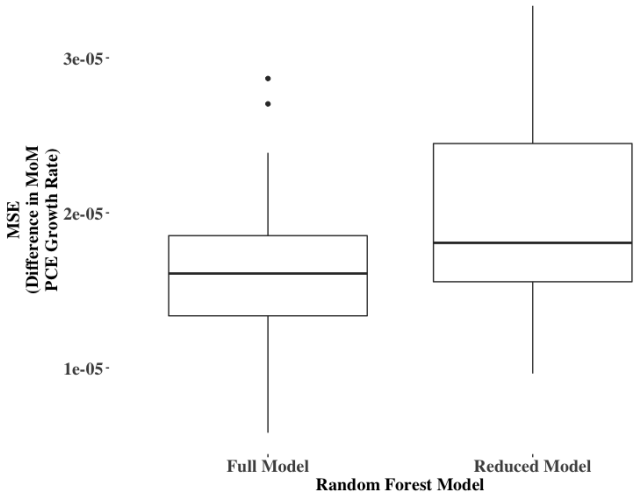


Fig. 11. Boxplot of MSEs of random forest models, $n = 50$.

Finally, we perform a Mann-Whitney U (MWU) test to compare the two sets of MSE metrics. The MWU is a non-parametric version of the two sample *t*-test, and computes the probability a statistic from one sample is not equal to that of another sample. This procedure results in a test statistic of 838 and a *p*-value of 0.002, thereby providing strong evidence that the MSEs of the two models are not equal.

6 Ethics

The United States has no comprehensive digital privacy laws. Companies in possession of personal data are under no obligation to inform individuals what they do with their data. They are not required to divulge what data they have collected, how they are using it, and with whom they are sharing it.

The absence of privacy laws and industry regulation does not absolve any data science practitioner of their responsibility to act ethically. The Association for Computing Machinery [18] and the Institute of Electrical and Electronics Engineers Software Society [19] both publish codes of ethics which any practitioner can look to for guidance. The Digital Ethics Labs at Oxford University is in the first stages of drafting ethical guidelines for the data science fields of machine learning and AI. [20] These codes have common themes of focusing on the public good, privacy, integrity, and human values. Ethical codes reinforce the notion that ethical practice is a voluntary endeavor and that adherence to an ethical code is incumbent upon each individual practitioner.

The speech we are analyzing was published in a very public forum so there was no expectation of privacy by the speaker. We were not explicitly given permission to extract latent information from the speech in the earnings call transcripts and use it beyond an evaluation of the associated company. In our analysis, we are collecting

economic information of a narrow scope and aggregating it to make inferences about macroeconomic conditions. Our use of sentiment derived from this speech does not cross any threshold of questionable use. Nor does our use of the sentiment violate the personal privacy of any one individual, group, or entity. However, out of an abundance of ethical caution, it must be assumed that our prediction of PCE could be an input to yet another predictive model which may have some effect. Any subsequent use of our predictions is beyond our control, though not beyond our consideration.

A crucial issue in NLP is the subjective nature of language. Word definitions can have subjectively different meanings to the speaker than to the listener. In our NLP application we used a lexicon containing more than 85,000 words. It is hard to estimate how this dictionary agrees with the vocabulary of the speakers. While we are predicting a very broad economic indicator that does not have a singular, direct effect on anyone to our knowledge, the possible bias in our lexicographic approach should not be overlooked.

The development and use of NLP techniques to analyze speech, in any form, has ethical implications in a broader social context. Speakers who are aware their speech will be analyzed tend to conform to normative speech patterns. [13] However, at present most individuals are unware of the capabilities of NLP techniques. A survey of international NLP practitioners showed that 91% of respondents indicated that they did not believe that the public was aware of the limits or possibilities of NLP tools. [14] The reinforcement of normative behaviors is an ethical issue that the NLP field will need to consider closely.

Securities laws regarding the disclosure of material information and insider trading exist to create a level playing field for investors. In the course of normal operations, executives of public companies possess pertinent knowledge that they have yet to disclose to the public. Using natural language processing to extract latent information from executive speech may subvert the intent of financial regulations preventing such disclosure.

Excepting specific laws to prevent the application of these techniques, it is up to investors using NLP to draw ethical lines when searching for profit. As we are predicting a broad macroeconomic indicator, we do not believe we have violated the spirit of any laws in our application.

In the paper *Ethical Fading: The Role of Self-Deception in Unethical Behavior* [15] the authors propose that unethical behavior happens as an incremental transition. These gradual changes occur because we frame trivial shifts in behavior as ethically neutral in an act of self-deception. In this understanding of ethical behavior, the first step away from the foundations of our value system might as well be the last. This “ethical fading” concept has significant implications for data science as its effects penetrate and reverberate through society.

9 Conclusions and Future Work

Models which include executive sentiment features derived from earnings call transcripts improve predictions of PCE. This analysis also leads us to conclude there are additional novel features to be mined from financial text sources which will further

improve predictions of PCE and other macroeconomic measures. NLP provides a powerful set of tools for quantifying data from financial texts. While opportunities for additional study along these lines abound, perhaps the most promising avenue would be the application of more sophisticated sentiment analysis methods which move beyond lexical databases.

The magnitude of prediction errors of machine learning models often decreases as the number of instances upon which they are able to train increases. Our research examined the relatively short time frame of January 2011 to June 2018, resulting in only 90 instances. Both the VAR and random forest models would benefit from more data. Obtaining more data from periods prior to those investigated here is difficult as the transcription of earnings calls is increasingly less common and the data sources become more disparate and less reliable when looking further back in time. However, future periods should provide at least equivalent coverage to the period examined here. Additionally, the period of analysis does not include any instances of macroeconomic recession, and are unlikely to generalize well to such an environment. Training the models investigated here on data obtained during such an environment may further improve the predictive power and robustness of the models, and thus warrants future study.

Machine learning continues to find applications in all data-centric disciplines, and macroeconomics is no exception. This trend is unlikely to abate, and all macroeconomic forecasting problems have the potential to benefit from the application of machine learning prediction techniques. In this paper, we have added to the body of evidence which shows, quite clearly, that machine learning algorithms are a highly effective set of tools for making macroeconomic predictions. While the random forest algorithm applied in this paper is effective, it is likely that other machine learning methods would also perform well on these data. One avenue of future work for this prediction problem would be to explore different machine learning and AI algorithms, as well as to conduct a deeper study of the optimal hyperparameter configurations of those algorithms.

References

1. Callen, T. (2017, July 29). Gross Domestic Product: An Economy's All. Retrieved August 11, 2018, from https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm
2. Bureau of Economic Analysis, Table 2.8.5. Personal Consumption Expenditures by Major Type of Product, Monthly. https://apps.bea.gov
3. Crawford, M., Davis, H., Jr., Brown, K., Johnson, E., & Aversa, J. (2018, October 30). Consumer Spending. Retrieved November 01, 2018, from https://www.bea.gov/data/consumer-spending/main
4. What We Do. (2013, June 10). Retrieved July 04, 2018, from https://www.sec.gov/Article/whatwedo.html U.S. Securities and Exchange Commission
5. Rochwarger, Jonathan P., and Miachika, Alexander. "Earnings Call Practices: How Does Your Company Compare to Others?" Lexology, McDermott Will & Emery, 26 Jan. 2015, www.lexology.com/library/detail.aspx?g=bc801101-b83d-4934-bd0f-71c01c0d3a04.
6. NIRI Earnings Process Practices Research Report. (2016). Alexandria, VA: National Investor Relations Institute.
7. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

8. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.
9. Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DC. Text mining for market prediction: A systematic review. Expert Systems with Applications. 2014 Nov 15;41(16):7653-70.
10. Wuthrich B, Cho V, Leung S, Permunetilleke D, Sankaran K, Zhang J, Lain W. Daily stock market forecast from textual web data. In IEEE International Conference On Systems Man And Cybernetics 1998 Oct 11 (Vol. 3, pp. 2720-2725). Institute Of Electrical Engineers Inc (IEEE).
11. Xing FZ, Cambria E, Welsch RE. Natural language based financial forecasting: a survey. Artificial Intelligence Review. 2018 Jun 1;50(1):49-73.
12. Kumar BS, Ravi V. A survey of the applications of text mining in financial domain. Knowledge-Based Systems. 2016 Dec 15;114:128-47.
13. Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 591-598).
14. Fort, K., & Couillault, A. (2016, May). Yes, we care! results of the ethics and natural language processing surveys. In international Language Resources and Evaluation Conference (LREC) 2016.
15. Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. Social justice research, 17(2), 223-236.
16. Howrey, E. P. (2001). The predictive power of the index of consumer sentiment. Brookings papers on economic activity, 2001(1), 175-207.
17. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.
18. ACM, C. M. (1992). ACM code of ethics and professional conduct. Code of Ethics.
19. Software Engineering Code of Ethics. (1999). Retrieved from https://www.computer.org/web/education/code-of-ethics IEEE Computer Society
20. King, T. and Floridi, L. (2018). ESIAI: Ethical and Social Implications of AI. [online] Digital Ethics Lab. Available at: https://digitalethicslab.oii.ox.ac.uk/esiai-ethical-and-social-implications-of-ai/ [Accessed 12 Dec. 2018].

Appendix: Sample Earnings Call Transcript

Excerpted from Apple Inc.'s second quarter 2018 earnings call transcript.

Operator
Your first question will come from Shannon Cross with Cross Research.
Shannon S. Cross - Cross Research LLC
Thank you very much. I wanted to ask about your thoughts on sort of iPhone and positioning now that we're a couple of quarters out from the launch of the iPhone X. Given the $1,000 price point, and it's clearly selling but there's been a lot of questions in the market about sustainability of that price point and how you're thinking about it as you look out sort of holistically across your lineup. So if you could talk a bit about what you're hearing from your customers and then I have a follow up. Thank you.
Timothy Donald Cook - Apple, Inc.
Sure. Shannon, it's Tim. As Luca mentioned earlier, our revenues are up 14% year-over-year on iPhone and that's a combination of single digit unit growth and ASP growth that is mainly driven by iPhone X. I think that our iPhone line shows that there's

a variety of different customers in a market that is as large as a smartphone market and so we're going to continue to provide different iPhones for folks to meet their needs.
On iPhone X specifically, I think it's important to maybe emphasize again one of the things I mentioned in my opening comments, that customers chose iPhone X more than any other iPhone each and every week in the March quarter, just as they did following its launch in the December quarter.
Also, since we split the line with the launch of iPhone 6 and 6 Plus back in 2014, this is the first cycle that we've ever had where the top of the line iPhone model has also been the most popular. And so with the customer set that Luca referenced as well, the 99%, the iPhone X is a beloved product. And so I think that it's one of those things where like a team wins the Super Bowl, maybe you want them to win by a few more points but it's a Super Bowl winner and that's how we feel about it. I could not be prouder of the product.
Shannon S. Cross - Cross Research LLC
Okay. Thank you. And then, Luca, can you talk a bit about working capital, specifically inventory which went up pretty significantly quarter-over-quarter? What's driving that and how are you thinking about, I mean, it's one of the uses of cash obviously, so how are you thinking about inventory and maybe working capital in general as you're going forward?
Luca Maestri - Apple, Inc.
Yeah, Shannon, you know that we've always generated significant amount of cash through working capital. We've got a negative cash conversion cycle and we plan to continue to have that. Our inventory level has gone up. It's just a temporary event. We have decided to make some purchasing decision, given current market conditions, and that should unwind over time.
Shannon S. Cross - Cross Research LLC
So that was essentially component purchases?
Luca Maestri - Apple, Inc.
Correct.
Nancy Paxton - Apple, Inc.
Thank you. Shannon.
Shannon S. Cross - Cross Research LLC
Okay. Thank you very much.
Nancy Paxton - Apple, Inc.
Can we have the next question, please?

Appendix: Python Code

The Python code for the analysis performed in this paper can be accessed at https://github.com/improvements-to-consumption-prediction/Improvements-to-Consumption-Prediction.git.