Essays in Applied Economics and Machine Learning

by

Ying-Kai Huang

BS in Mechanical Engineering, National Taiwan University, 2012

MA in Economics, University of Pittsburgh, 2018

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Ying-Kai Huang

It was defended on

June 14th, 2021

and approved by

Douglas Hanley, Department of Economics, University of Pittsburgh

Satish Iyengar, Department of Statistics, University of Pittsburgh

Arie Beresteanu, Department of Economics, University of Pittsburgh

Jean-Francois Richard, Department of Economics, University of Pittsburgh

Essays in Applied Economics and Machine Learning

Ying-Kai Huang, PhD

University of Pittsburgh, 2021

This dissertation consists of three chapters in applied behavioral economics and machine learning applications in economics.

The first chapter studies how reference-dependent utilities influence people's behaviors on crowd-sourced review websites and cause attribution bias. Using data from Yelp, I tested how potential disappointments may affect customers' reviews by applying a regression discontinuity design to control for unobserved factors that may also simultaneously influence ratings. This chapter links to an emerging literature of attribution bias in economics and provides empirical evidence and implications of attribution bias on online reputation systems.

The second chapter extends the work of first study and explores attribution bias when both reference dependence and state dependence are possible to appear. I specifically use the scenario of special occasions to test two leading theories of attribution bias empirically. The empirical results can be explained by one theory of attribution bias where people have higher expectations about restaurants on special occasions and then misattribute their disappointments to the qualities of the restaurants. From the connection between our empirical analyses and theories of attribution bias, this chapter provides another piece of evidence of how attribution bias influences people's perceptions and behaviors.

The third chapter connects machine learning with financial forecasting. I construct a model with recurrent neural networks and focus on the point forecasting of the yield curve to explore the possibility of having better forecasts for the term structure. While allowing similar interpretation as previous econometric methods, the neural network model in this paper shows better forecasting accuracy.

Table of Contents

List of Tables

List of Figures

Preface

Firstly, I want to thank God for leading me to the program and helping me complete this dissertation. The Lord is my shepherd; I shall not want.

I am deeply indebted to my advisor, Douglas Hanley, who gave me the courage to continue my studies when I was lost and frustrated with research. His support and guidance has led me through the hard time and made me a better researcher and a better man. I am also fortunate to have Satish Iyengar as my co-chair. I am extremely grateful for his extra effort on me, and the discussions we had about my research are the unforgettable memories in my PhD time. I would also like to thank Arie Beresteanu and Jean-François Richard for their research advice and service on the committee. Without them, it is not possible to complete this dissertation. Special thanks to Daniele Coen-Pirani, Brian Deutsch, Ian Morrall, Domonkos Vamossy, David Min-Heng Wang, Neil Ni, Lucy Zhang and many other graduate students in the Economics Department at University of Pittsburgh for their friendship and support.

I also want to thank my parents for giving me the freedom to chase my dream and alway supporting me. Finally I would like to thank my wife and co-author, Shuyan, for inspiring me with research ideas, grammar checks and her unconditional love. I love her to the moon and back.

1.0 Attribution Bias on Online Reputation Systems

Consumers benefit from reading ratings online before making their purchases, yet this information aggregation process may have some potential problems that was not previously credited in the literature. Through an empirical approach, I showed how people could review businesses inconsistently when their expectations are formed by ratings on crowd-sourced review websites. Using data from Yelp, I tested how potential disappointments may affect customers' reviews by applying a regression discontinuity design to control for unobserved factors that may also simultaneously influence ratings. In addition, I developed a model illustrating rating behaviors with reference-dependent utilities to establish testable hypotheses and showed that comparisons between their true experience and expectation, when consumers write their reviews, could impede their assessment of businesses' qualities and cause attribution bias. After carefully excluding confounding variables, my results support the hypothesis that consumers have attribution bias when they write reviews. Several robustness checks support these findings and shed further light onto this example of attribution bias. This paper links to an emerging literature of attribution bias in economics and provides empirical evidence and implications of attribution bias on online reputation systems.

1.1 Introduction

Consumer-reviewed websites have increasingly attracted attention over the past decades. More and more people use these websites as guides when it comes to searching for restaurants, hotels, movies. . . etc. These websites create platforms for people to gather information at a low cost and decrease the level of information asymmetry. However, this process of information aggregation may be causing a variety of problems. Among drawbacks of existing online reputation systems, reference-dependent behaviors are one important factor that was not widely-recognized and can undermine the credibility of review websites.

Reference dependence means that people's experiences are affected by expectations they have in their minds. Imagine Bill was heading to a restaurant for dinner. Before he left, he had checked reviews on Yelp and found out the restaurant's rating was 4 out of 5 stars. Bill probably would expect good food and nice service. However, if the dining experience turned out to be a 3.5-star level to him, would he feel as though the experience was poorer than the case which he did not check the reviews beforehand? The answer supported by evidence in this paper is yes. The comparison between his true experience and expectation made the dining experience less enjoyable. With this psychological effect, when Bill is asked to review the restaurant, he would tend to give a lower rating than he would have, if he did not have the reference. This phenomenon is called *attribution bias* because Bill misattributed his psychological loss to the quality of the restaurant. Since an ideal rating system should truthfully reflect the quality of businesses, attribution bias is one important drawback of crowd-sourced sharing platforms that should be carefully studied and explored. This paper is the first to discover attribution bias in online rating systems, providing an important contribution of psychological bias in the e-commerce literature.

To discover attribution bias on consumer-reviewed websites, this paper studies review data from Yelp and takes advantage of special structures of the platform. In most crowd-sourced review forums, reviews are presented in numerical and/or graphical forms. In particular, ratings on Yelp are presented in stars and users are only allowed to use full stars from 1 to 5 in rating a business. For instance, a reviewer can give 3 stars but cannot give 3.5 stars on the website. In addition to individual user ratings, Yelp also calculates average ratings in half-star increments for businesses and shows the average ratings as the most apparent measure of the businesses on the website. Since units of average ratings are in half stars, the rating system has to round average ratings of businesses into the nearest half-star units. For example, an average rating of 3.75 would be rounded up to 4 stars and a 3.74 would be rounded down to 3.5. This rounding mechanism on Yelp creates discontinuities for businesses with similar ratings and creates potential psychological loss for consumers. Applying this concept to the story of Bill, if the restaurant had an average of 3.75 stars, Bill would see the average rating as 4 stars on the website and the 0.25 difference between the true average and the rating he saw would become his disappointment and could potentially influence his experience. From Yelp data, if Bill wrote his review after his dinner, I could further observe whether his review showed attribution bias. This idea can be generalized to other consumers as well. By using the data from Yelp and its discontinuities, I could check whether attribution bias affected people's ratings and caused a problem for the rating system. The rounding feature on Yelp also lends us a good way to control for many unobserved effects. Since rounding thresholds on Yelp are exogenous, the discontinuities in the data gave a treatment that was as good as random and created a natural experiment that excluded many endogenous factors. By utilizing the data features, this paper applies a regression discontinuity design as an identification strategy to detect attribution bias and control for possible confounding factors.

Through the empirical analysis, this paper contributes to the reference dependence literature in behavioral economics. Reference dependence has been introduced to economics by [48]. While scholars have gradually realize the importance of reference dependence, we may still neglect its existence in many fields. More evidence of this effect could help us better understand when this phenomenon could be important. Previous work has been done, for example, on how a small tax affects peoples' behaviors in using plastic bags [41] and how loss aversion could influence the asking price in the housing market [31]. This paper, on the other hand, identifies the existence of reference dependence when people write their reviews on crowd-sourced review websites and helps us understand its importance in online reputation systems. Furthermore, this study takes a step beyond reference dependence by linking its findings with attribution bias. Reference dependence only explains how people's behaviors are affected by their expectations in mind. Attribution bias further describes how reference dependence can distort peoples' perceptions and affect how people learn about qualities of products, usefulness of information, and values of many other things. This paper connects to attribution bias by studying how attribution bias affects consumers' learning of businesses' qualities and further influences their reviews.

Learning with attribution bias is an emerging research topic in behavioral economics. In the literature, some theoretical models have been proposed to explain how learnings can be misled by attribution bias. [30] formalized attribution bias into a theoretical model and use the model to describe how consumers can misattribute their gain-loss utilities when their experiences deviate from their expectations. For experimental and field studies, [36] found that consumers could misattribute temporary states, such as weather or thirst level to the stable qualities of consumption goods through experiments and surveys. [10] conducted laboratory and online experiments to demonstrate how positive and negative surprises could affect workers' perceptions of effort costs and influence their willingness to work. Built on the behavioral evidence supported by the above-mentioned research, this paper further consolidates the existence and importance of attribution bias by providing the first empirical evidence from observational data in this literature.

Another contribution of this paper is in the industrial economics and online-review literature. One area which this study helps explain is people's motivation for rating businesses. In [54], an experiment was conducted to demonstrate when buyers know true qualities of sellers after their consumption; buyers tend to give higher ratings for sellers who exceeded their expectations and revenge to those with lower qualities. This paper connects the model in [54] to the model of attribution bias in [30] and create a new theoretical framework showing that attribution bias could be one cause of buyers' reactions in the real world. Previous works have also shown that ratings on Internet review forums could possibly affect businesses' revenue. [1] demonstrated that a half-star increase in Yelp rating decreases the online reservation availability by 19 percent and increases restaurants' revenue significantly. [62] estimated that a one-star increase in rating can increase revenues by 5-9 percent for restaurants without chain affiliation. [84] showed how higher ratings on websites could possibly hurt sales. [17] presented how patients' reviews on Yelp can increase the revenue of clinics and benefit other patients by informing them about the medical service qualities of physicians. Although many research had been done to understand the advantages and possible drawback of online-review system, attribution bias in consumer rating systems was not credited in the literature previously. Combined with the above findings, this paper provides another perspective on how attribution bias can be a channel affecting the revenues of businesses and the welfares of consumers.

The rest of the paper proceeds as follows. In the next section, a theoretical framework is provided to establish testable hypotheses. A two-period model is proposed to fit into the learning and rating environment we usually see in the real world. Section 3 is data summary and identification strategies. In section 4, I present the main estimation results and robustness checks. In the last section, I include discussion about implications of attribution bias and possible future work.

1.2 Theoretical Framework

In this section, I develop a model illustrating rating behaviors and generate propositions based on the model. The two-period model characterizes how consumers learn about the quality of businesses in the first period and rate businesses after their consumption experience in the second period.

1.2.1 The Model

The model I propose is an extension from [30] and [54].

1.2.1.1 First Period

In first period, a consumer i learns the quality of a business using his recent dining experience. The utility from the business takes the following form:

$$u_i(q|r_i) = \underbrace{v_i(q)}_{\text{Consumption utility}} + \underbrace{\mu_i \; n_i(v_i(q)|r_i)}_{\text{Gain-loss utility}} \tag{1}$$

$$n_i(v_i(q)|r_i) = \begin{cases} f_i(|v_i(q) - v_i(r_i)|) \text{ if } v_i(q) \geq v_i(r_i) \\ \lambda_i f_i(|v_i(q) - v_i(r_i)|) \text{ if } v_i(q) < v_i(r_i) \end{cases} \tag{2}$$

$q \in [1, 5]$ is the quality of the business, r_i is the reference point and μ_i is the coefficient governing the impact of gains and losses. $v_i(q)$ is the utility from the dining experience and it is increasing in quality. $f_i(x)$ is a concave function increasing in x with $f_i(0) = 0$ and $\lambda_i \leq -1$ captures loss aversion. The consumer then tries to figure out the quality of the businsee from her experience. If she is a rational agent, she can exclude the gain-loss utility and ascertain the quality of the business. However, if she cannot fully eliminate the influence from gain and loss, she may misattribute her bad experience relative to her reference point to the true quality and undervalue the business, and vice versa. To be more specific, consider an example that consumer i goes to a business with quality q and totally ignores the gain-loss utility. In such a case, the consumer experiences an utility level of u^* and

$$u^* = v_i(q) + \mu_i \; n_i(v_i(q)|r_i) \tag{3}$$

However, since the consumer ignores the gain-loss utility, she will believe $u^* = v_i(q)$ and form an estimate $\hat{q}$ of the quality of the business based on consumption utility only. Thus, when there is a bad experience, the consumer will include the loss utility when she considers the quality of the business without notice and will undervalue the business since $\hat{q} < q$. This phenomenon is the so-called attribution bias.

1.2.1.2 Second Period

After learning the quality of the business, consumer i decides whether to rate it online. Rating is costly and only if the benefit from rating a restaurant is greater than the cost, will consumer i be willing to write a review. Moreover, expressing the true experience is easier than giving other kinds of ratings since coming up with imaginary reasons always takes more effort. In my model, I also assume a general consumer cares about both businesses and other customers. When consumer i is considering how to rate a business, if consumer i's consumption experience exceeds her expectations, she has an altruistic feeling for the business and gives a higher rating. On the other hand, if her experience with the business is worse than her expectation, she will take a vengeful action and give a lower rating. When it comes to the concern about other consumers, I assume consumer i's utility takes a warm glow form. That is when considering whether reviewing businesses will help others, she cares about whether she writes a review or not but does not care about how useful or precise the review is. To sum up, the maximization problem of consumer i in the second period is:

$$\max_{a_i} \; U_i(\hat{q}_i, r_i, a_i) = \max_{a_i} \; \omega_i(\hat{q}_i - r_i)U_s(a_i) + \beta_i \sum_{j}^{n} I_a - I_a c_i(|a_i - \hat{q}_i|) \tag{4}$$

r_i and $\hat{q}_i$ are the reference point and the quality she learns from the first period. $a_i \in [1, 5]$ is the rating she can give. $U_s(a)$ is the utility of the sellers and it is increasing in a. ω_i shows how much she cares about whether a business meets her expectation and it can influence the strength of her reciprocal or vengeful action. I_a is the indicator function equals 1 if consumer i writes her review. $\beta_i \sum_j^n I_a$ is the utility from helping other consumers and it is a constant once consumer i decides to rate. This utility function captures the warm glow of consumer i. The warm-glow assumption simplifies the analysis and its comparative statics will not change if I replace the assumption with the sum of other consumers' utility as long as their utility functions are concave and they prefer a precise review. $c_i(x)$ is the cost of writing reviews and it is a convex function with minimum at $x = 0$. Consumer i takes $\hat{q}_i$ and r_i into account, decides whether to rate or not and what is the a_i^* she wants to give and maximize $U_i(\hat{q}_i, r_i, a_i)$.

1.2.2 Propositions

After analyzing the model proposed above, we can conclude some comparative statics of the model to the following propositions.

Proposition 1. *When there is attribution bias and* $q < r_i (q > r_i)$*, the perceived quality,* $\hat{q}_i < q$ *(*$\hat{q}_i > q$*).*

Proposition 1 gives us a scenario to test the existence of attribution bias. If two groups of reviewers both have attribution bias and different reference points, the perceived qualities of them will be different.

Proposition 2. *When consumer* i *decides to rate and* $\hat{q}_i < r_i$ *(*$\hat{q}_i > r_i$*) and if there is attribution bias and* $\omega_i = 0$ *then* $a_i^* = \hat{q}_i$ *(*$a_i^* = \hat{q}_i$*).*

From proposition 2, if we can find an environment where $\omega_i = 0$ and two groups of reviewers have different reference points and attribution bias, then we should see a difference in their rating behaviors. Furthermore, the difference demonstrates the impact of attribution bias.

1.3 Data Source and Identification Strategy

1.3.1 Data Source

Established as a restaurant review forum in 2004, Yelp.com is now the largest crowd-sourced review website in North America for all businesses, including medical services, home improvement and many other industries. In 2019, Yelp has 36 million unique users and has accumulated 205 million reviews. Yelp not only provides its users with valuable information about businesses but also creates a convenient platform to share their own experience and makes the information aggregation process easier than ever. The company also provides researchers a great data source to understand its users' behaviors by holding many rounds of data challenges. I seized this opportunity provided by Yelp and conducted empirical analysis in this paper based on the released data from Yelp data challenge round 13. The data includes all recommended reviews in 11 metropolitan areas around the world. The 11 cities are Edinburgh, Stuttgart, Montreal, Toronto, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, and Cleveland from October 2004 to November 2018.

1.3.2 Identification Strategy

The ratings on Yelp are in half-star increments. This means the stars showed on the website are rounded to the closest half-star units. For example, if the underlying true average of a restaurant is 3.75, users will see 4 stars, and if the true average is 3.74, users will see 3.5 stars. By utilizing this rounding feature, I assumed that a true quality of a business is its true average and consumers form their expected quality of the business based on the rounded stars they see on Yelp. When the true average of the business is inconsistent with the stars users perceive, the difference between the true quality and consumers' expectation could cause a psychological loss or gain for them. For instance, when the true average of a business is 3.75 and the star on Yelp is 4, the gap between the business's true quality and users' 4-star expectation would be a loss. Since these gaps in ratings can be reasonably modeled as exogenous random shocks in small windows, the environment allows me to use the regression discontinuity method to test the effect of reference dependence and further explore if it is caused by attribution bias.

Regression discontinuity method was first proposed in [79]. The two psychologists use regression discontinuity in their seminal paper to test whether National Merit Scholarship awards affect the career choice of college students. The statistical theories, limitations and applications behind regression discontinuity design are provided by [57] in details. In my analysis, I focused on an environment where the true averages of businesses are close to the rounding thresholds. This choice allowed me to utilize discontinuities in the rounded average ratings and apply a regression discontinuity design. One important assumption behind this design is that the only difference for businesses on different sides of the rounding thresholds is the ratings which consumers saw on Yelp. I argue this is a reasonable assumption because other things related to businesses' quality should be truthfully reflected by their true averages and do not change significantly with the rounding thresholds. This data characteristic gives me a natural experiment to test the treatment effect of reference dependence since the only way the discontinuities can influence consumers' experience is through manipulating their expectations of the businesses before their visits. To estimate the effect of reference-dependent behaviors, I run the following ordinary least square regression:

$$star_i = \alpha_0 + \alpha_1 RD_i + \alpha_2 \text{diff}_i + \lambda X_i + \epsilon_i \tag{5}$$

$$RD_i = \begin{cases} 1 \text{ if } \text{diff}_i > 0 \\ 0 \text{ if } \text{diff}_i \leq 0 \end{cases}$$

Here $star_i$ represents the rating consumer i posted after she visited a business. diff_i is the distance between the true average and the rounding threshold before consumer i went to the business. RD_i is an indicator function showing whether the true average of the business was on the right-hand side or left-hand side of the threshold. When $RD_i > 0$, the true average was on the right-hand side of the threshold and the star on Yelp was rounded up to the closest half unit. Thus, when $RD_i = 1$, the gap between the true average and the star on Yelp caused a potential loss for consumer i. Similarly, when $RD_i = 0$, there was a potential gain for consumer i. By estimating the coefficient of RD_i, we can quantify how different reference points affect consumers' rating behaviors. X_i represents control variables and ϵ_i is the error term. In Yelp's data, $star_i$ is actually a discrete variable of an integer value ranging from 1 to 5. Thus, I also used ordered Probit regression in my analysis to model the discrete choices of reviewers.

A closer look on the equation (5) helps us understand how the proposition 1 and 2 in the theoretical framework can be tested. As a result of the proposition 1, when there is attribution bias, we expect the perceived qualities $\hat{q}_i$ would be different if users were on different sides of the rounding thresholds. From the proposition 2 and the difference in perceived qualities, we also expect the rating actions a_i^* would be different. This variation allows us to use the regression to detect the reference-dependent behaviors of Yelp users. If the estimation of equation (5) shows a significant negative effect of RD, it will provide an supporting evidence of attribution bias in reviewers' ratings.

1.3.3 Data Selection

In my main analysis, I only included reviews for businesses with more than 100 ratings on Yelp at the time they were reviewed. This restriction was made to ensure my analysis was conducted on relatively stable businesses. To apply the regression discontinuity design, I also set up a 0.005-unit bandwidth to focus my analysis in small windows. Leveraging such a large dataset, I am able to choose such a fine bandwidth to help ensure that using a linear functional form to approximate the data trend is suitable in my analysis. This assumption will be further tested in the following section of bandwidth choice. In addition, since one of the main goals of this paper is to prove the existence of attribution bias, I only checked ratings from consumers who reviewed the businesses the first time. These restrictions will be relaxed and tested in the following sections and the main implications of this paper do not change much. The original data does not have the average stars of each business that reviewers could see in different time periods. Thus, I calculated the average stars for all businesses at every time point from the data. With the restrictions and the adjustments mentioned above, I present the summary statistics of my data in the appendix.

1.4 Empirical Analysis and Estimation Results

In this section, I show the estimation results of the previous specified regression. All standard errors in my estimations are calculated with the method of [85] to account for potential heteroscedasticity. In addition, I also provide several placebo tests to validate my regression results.

1.4.1 Reference Dependence in Rating Behaviors

To begin with, I start my analysis by providing a graph to visualize the discontinuity in the data.

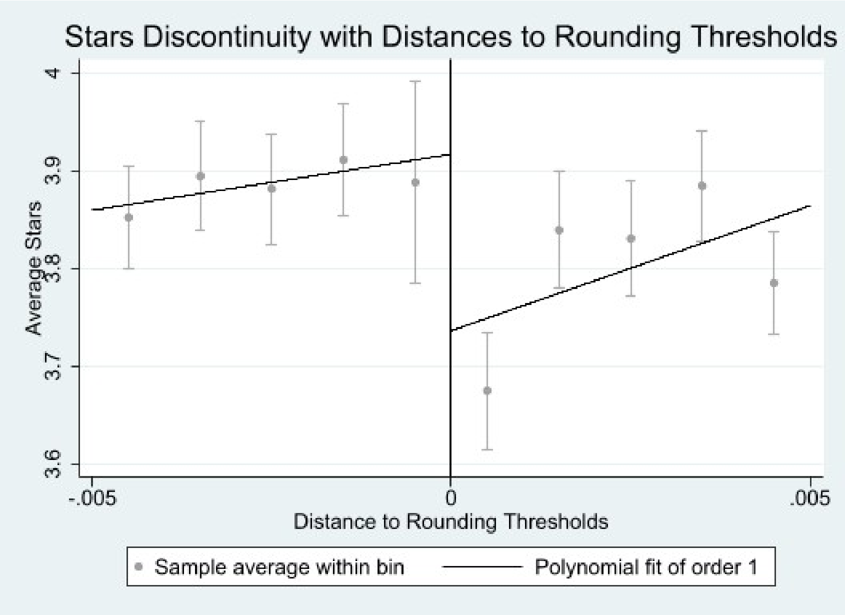


Figure 1: Stars Discontinuity with Distances to Rounding Thresholds

In this figure, the 0.005 unit is chosen to be the bandwidth of the figure. I further divided the diff axis into ten parts. Each part has a length of 0.001 unit and can be considered as one small bin on the graph. Within the bins, data points are pulled together. For example, if one data point is 0.00025 unit to the left-hand side of 0, the data point will be assigned to the bin covering -0.001 to 0. Each red dot represents an average of one bin and black lines are the visualization of the estimation results of the equation (5).

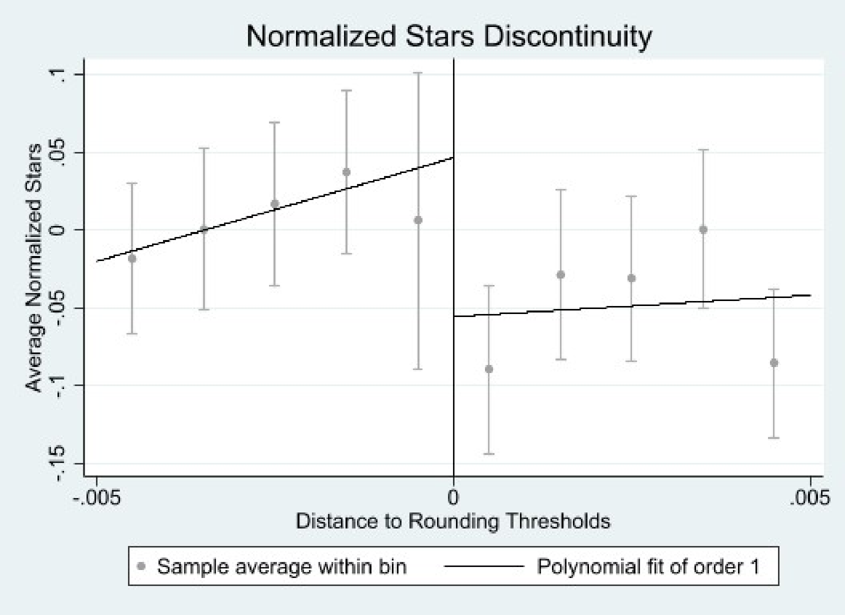


Figure 2: Normalized Stars Discontinuity with Distances to Rounding Thresholds

From figure 1, we can see there is a clear discontinuity before and after the threshold. In figure 2, I normalized the ratings by subtracting the stars with their closest rounding thresholds. For instance, when a business has a true average of 3.23, the closest rounding threshold is 3.25 and the ratings given by users is subtracted by 3.25. In the normalized graph, the average of the normalized stars before the thresholds are higher than 0 and the average of normalized star are lower than 0 after the thresholds. Moreover, the pattern of the graph is similar to the one without the normalization. To check whether this discontinuity also appears after controlling other possible covariates, I present the following regression results.

In the first column of table 1, I estimate equation (5) without any control variables. The result shows the effect of RD is significant. From my previous argument, we can also interpret this result to mean that there are significant differences in consumers' rating behaviors because consumers had different reference points. One possible concern about the result in column 1 is that people may react very differently when they go to businesses in different star ranges. For example, the rating behavior for a person going to a 4.5-star

Table 1: Regression Results of Reference-Dependent Behaviors

	(1)	(2)	(3)	(4)
	Reviewer's Rating	Reviewer's Rating	Reviewer's Rating	Reviewer's Rating
RD	-0.1955***	-0.5892***	-0.5846***	-0.5847***
	(0.041)	(0.039)	(0.039)	(0.039)
diff	20.6772***	7.4264	7.1154	7.1672
	(6.646)	(6.060)	(6.077)	(6.085)
intercept	3.9439***	1.3394***	1.5683***	1.7292**
	(0.024)	(0.247)	(0.471)	(0.717)
Observations	20897	20897	20897	20897

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the estimation without any controls

(2) adds fixed effects for different rating ranges

(3) adds city fixed effects and fixed effects for different rating ranges

(4) adds above fixed effects and year fixed effects

restaurant is very likely to be different from the situation when she goes to a 3-star one. In column 2, I include dummy variables for star ranges as a way to control for this phenomenon. The result in column 2 alleviates this concern since the estimate of RD is again significant and its size is even larger. Column 2 shows how reference points can have a non-negligible impact on consumers' behaviors. After controlling the star-range effect, when consumers have higher expectations for businesses, their ratings can be a half star lower compared to the case when they have low expectations for similar businesses. In column 3 and 4, more fixed effects such as city fixed effects and year fixed effects are controlled in equation (5). The results in these two columns are similar to column 2 and provide evidence of users' reference-dependent behaviors.

When we carefully search for other possible explanations of the regression results in table 1, some concerns may arise. It is likely the results were caused by other mechanical reasons in the data and were not related to reference dependence. One potential explanation is mean reversion. Mean reversion is usually used to describe fluctuations in stock markets. A stock going up one day has a high chance to go down the next day. Prices of stocks are very random and do not really have a pattern. If this phenomenon also happened in the Yelp data, the rating differences caused by RD should not be considered as reference-dependent behaviors. To answer these concerns, I conducted a placebo test by creating imaginary thresholds in the data. In my placebo test, I chose thresholds that would not cause discontinuities in the stars consumers saw on Yelp. For instance, one threshold is 3.5, and no matter whether the true average was slightly below or above 3.5, consumers would see a 3.5 star. In that case, there is no potential gain or loss when consumers go to the business. Thus, if we cannot see any effect for those imaginary thresholds, the concerns about other mechanical factors should be mitigated. With this placebo test, I estimated the equation (5) on reviews which were written when the true average of the businesses were close to the stars consumers saw on Yelp. The result is presented in table 2. In column 1, we can see that the effect of RD for those imaginary thresholds is almost zero and insignificant. This result reassures us that the effect of RD should come from peoples' reference-dependent behaviors rather than from random incidents. Other placebo tests using imaginary thresholds with more controlled variables are provided in column 2, 3 and 4. The effect of RD is also not significant in these regressions.

Table 2: Placebo Test of Reference-Dependent Behaviors

	(1) Reviewer's Rating	(2) Reviewer's Rating	(3) Reviewer's Rating	(4) Reviewer's Rating
RD	-0.0056	-0.0230	-0.0202	-0.0195
	(0.029)	(0.028)	(0.028)	(0.028)
diff	3.5165	5.3597	4.5994	4.4960
	(4.831)	(4.556)	(4.604)	(4.618)
intercept	3.7709***	1.4763***	1.8096***	1.6092**
	(0.012)	(0.073)	(0.270)	(0.718)
Observations	30955	30955	30955	30955

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the estimation without any controls

(2) adds fixed effects for different rating ranges

(3) adds city fixed effects and fixed effects for different rating ranges

(4) adds above fixed effects and year fixed effects

To study these results more carefully, I also conducted ordered Probit regressions since users' choices are discrete. The results for the discrete choice model are provided in the appendix and conclusions from it are similar to what we have in table 1 and 2. From the results in the OLS model and the ordered Probit model, I provided evidence of Yelp users' reference-dependent behaviors with several placebo tests and showed that my findings are robust.

1.4.2 Revenge or Attribution Bias?

So far, the regression results in this paper provide strong evidence to support the existence of reference-dependent behaviors. However, it is not clear if Yelp users' reference dependence was caused by attribution bias or they simply just wanted to exact revenge, to be reciprocal or to correct the ratings when they noticed there were gaps between the qualities of the businesses and their ratings on Yelp. To answer this question, I proposed another regression to detect if the discontinuities in the data increase users' tendency of giving extremely low ratings. In this analysis, I run the following logistic regression:

$$P(RevengeReciprocal_i = 1|RD_i, X_i, \text{diff}_i) = F(\alpha_0 + \alpha_1 RD_i + \alpha_2 \text{diff}_i + \lambda X_i + \epsilon_i) \tag{6}$$

Here,

$$F\left(x\right) = \frac{1}{1+e^{-x}}$$

$$RevengeReciprocal_i = \begin{cases} 1 & \text{if } |\text{true average} - \text{star}_i| > 2.5 \\ 0 & \text{otherwise} \end{cases}$$

In this regression, $RevengeReciprocal_i$ captures if user i wants to exact revenge or to be reciprocal to businesses that had qualities different from her expectation. I assumed the reviewer would give an extremely low or high rating in this case. A revenge is defined by giving a rating that is 2.5 stars lower than the true average, and reciprocal is defined as giving a 2.5-star higher rating. The equation (6) allows me to estimate if RD affects the probability of seeing extreme ratings. The results are shown in table 2.

In the first column of table 3, the regression result without any controls shows a marginally significant effect of RD on the probability of revenge or reciprocal. After adding some rating range fixed effects, the effect of RD becomes stronger in column 2. Column 3 and 4 separate revenge and reciprocal, and both results show significant effect of RD on revenge or reciprocal probabilities. These regressions confirm the concern about factors confounding with attribution bias. To alleviate this concern, I utilize one special feature on Yelp's rating system.

Table 3: Regression Results of Probability of Revenge or Reciprocal

	(1)	(2)	(3)	(4)
	P(Revenge or Reciprocal)	P(Revenge or Reciprocal)	P(Revenge)	P(Reciprocal)
RD	0.1450*	0.6735***	0.8153***	-0.5165**
	(0.081)	(0.086)	(0.110)	(0.205)
diff	-20.5455	-9.624	-1.7157	-2.0198
	(13.020)	(13.619)	(17.246)	(32.618)
intercept	-1.7192***	-2.7430***	1.3394***	-2.7510***
	(0.047)	(1.034)	(0.247)	(1.056)
Controls	No	Yes	Yes	Yes
Observations	20897	20897	20897	20897

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the logit regression on vengeful and reciprocal actions without any controls

(2) shows the regression on vengeful and reciprocal action with fixed effects for different rating ranges

(3) shows the logit regression on revenge probability with fixed effects

(4) is the logit regression on reciprocal probability with fixed effects

Yelp has a special rewarding system for dedicated users and awards the best of them elite status. An elite can enjoy special events held by Yelp and interacts closely with other elites in the community. Attaining the level is like joining a private club and this rewarding system creates an additional incentive for users. To become an elite, a user needs to use her real name, have a clear photo of herself and post quality reviews frequently. Yelp also allows other users to evaluate how useful a review is. The usefulness of reviews is another important factor that will determine whether a user can become a member of Yelp's elite squad. This selection criterion for elites gives me a way to find an environment where users care much more about the usefulness and accuracy of their reviews than rating emotionally. Combining this characteristic of elites with the proposition 2 in my theoretical framework, I argue that $\omega_i = 0$ when I focus my analysis on Yelp elites. Therefore, the effect of RD excludes the impact from vengeful or reciprocal actions and can serve as evidence of attribution bias.

To demonstrate that elite members try their best to provide precise information about businesses, I estimate a variant of the equation (6) by considering whether a user had become an elite before she wrote her reviews. The regression which I extend from the equation (6) becomes:

$$\begin{aligned} &P(RevengeReciprocal_i = 1 | RD_i, X_i, \text{diff}_i) = \\ &\quad F(\alpha_0 + \alpha_1 RD_i + \alpha_2 \text{diff}_i + \alpha_3 elite_i + \alpha_4 RD_i \times elite_i + \lambda X_i + \epsilon_i) \end{aligned} \tag{7}$$

The regression results are shown in table 4. In the first column of table 4, we can see a clear negative effect of *elite*, which means when a user has become an elite, the probability of giving extreme ratings is much lower. This result supports our assumption that elite users do not rate emotionally. Furthermore, the interaction term of RD and *elite* is not significant in column 1, which suggests the discontinuities in the rating system do not have significant effect on elite users for the probability of giving extreme ratings. In column 2 to column 4 of table 4, I provide a more careful analysis by adding control variables or separate revenge and reciprocal actions. The results still support my assumption about elite users.

Table 4: Regression Results of Probability of Revenge or Reciprocal

	(1)	(2)	(3)	(4)
	P(Revenge or Reciprocal)	P(Revenge or Reciprocal)	P(Revenge)	P(Reciprocal)
RD	0.5326***	0.7366***	0.8519***	-0.7951*
	(0.106)	(0.111)	(0.112)	(0.441)
elite	-1.5803***	-2.2880	-1.2905	-2.4156
	(0.186)	(8.260)	(61.820)	(9.243)
$RD \times elite$	0.2570	0.0118	-0.0024	-5.0296
	(0.221)	(0.439)	(0.454)	(12.989)
Controls	No	Yes	Yes	Yes
Observations	20897	20897	20897	20897

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the logit regression on vengeful and reciprocal actions without any controls

(2) shows the regression on vengeful and reciprocal action with fixed effects for different rating ranges

(3) shows the logit regression on revenge probability with fixed effects

(4) is the logit regression on reciprocal probability with fixed effects

After testing the assumption of $\omega_i = 0$ in proposition 2, I estimated equation (5) again with elite users only. The regression results are in table 5.

In table 5, I present the estimation with elite users only in the first column. The corresponding result of non-elite users is in column 3. In column 1, we can see the significant negative effect of RD on reviewers' ratings and the size is non-negligible. From my estimation, when Yelp users had different reference points before they reviewed the same businesses, the ratings they gave could differ by a half star on average. Compared with the result in column 3, in which I included non-elite users in the analysis, we can see RD has a smaller effect when we only consider elite users. This result consolidates my assumption about Yelp elites because after removing non-elite users, we expect to see a smaller coefficient on RD since elites usually rate more objectively. Columns 2 and 4 provide placebo tests for columns 1 and 3. In these columns, the effects of RD are again insignificant and these results support that the effect of RD is from attribution bias.

Table 5: Attribution Bias of Elite Users

	(1) Reviewer's Rating	(2) Reviewer's Rating	(3) Reviewer's Rating	(4) Reviewer's Rating
RD	-0.4146***	-0.0372	-0.6281***	-0.0199
	(0.016)	(0.029)	(0.045)	(0.033)
diff	-2.5688	9.1575	10.1308	4.8003
	(10.383)	(7.699)	(7.062)	(5.324)
intercept	0.9919***	2.6146***	1.3945***	1.4239***
	(0.033)	(0.455)	(0.280)	(0.072)
Controls	Yes	Yes	Yes	Yes
Observations	3924	5888	16973	25067

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) adds fixed effects for different rating ranges and focuses on Yelp elites

(2) is a placebo test with imaginary rounding thresholds for (1)

(3) adds fixed effects for different rating ranges and focuses on non-elite users

(4) is a placebo test with imaginary rounding thresholds for (3)

1.4.3 Bandwidth Choice of Regression Discontinuity Design

Another possible concern about my regression results is the choice of bandwidth in my regression discontinuity design. It is possible that the above results only hold for some specific bandwidths. To answer this question, I followed [47] and use their nonparametric method to choose the optimal bandwidth. The discontinuity visualization is provided in figure 3 and the set of regression results is presented in table 6.

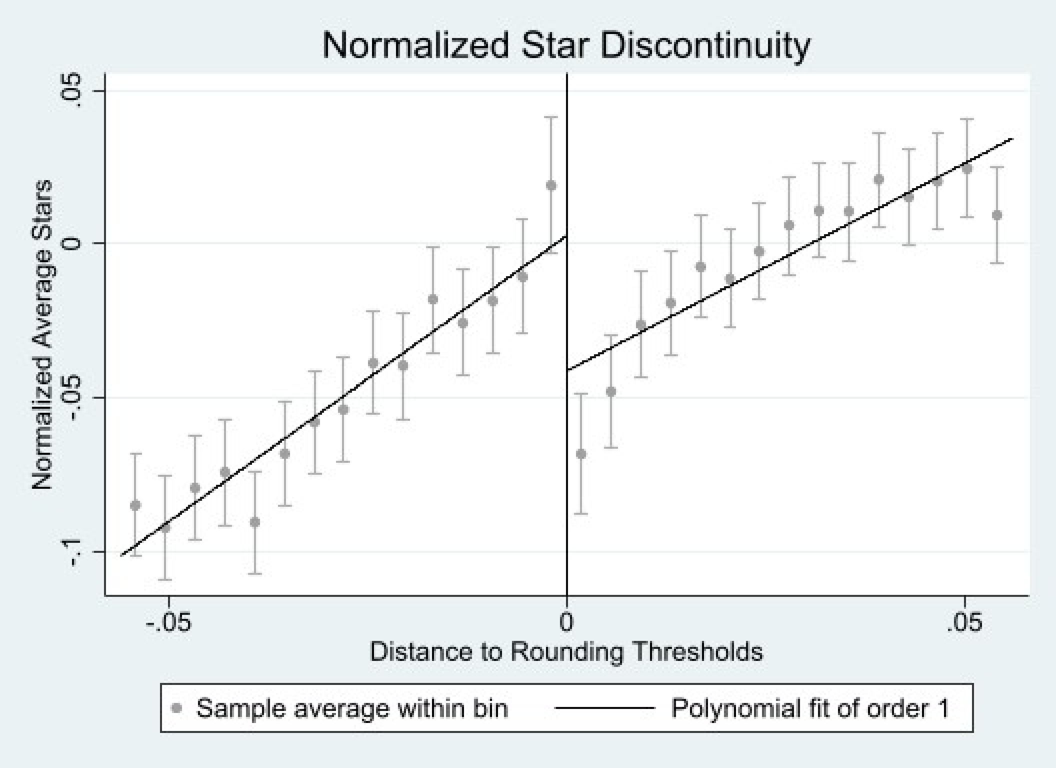


Figure 3: RD Plot with Optimal Bandwidth

Table 6: Attribution Bias with the Optimal Bandwidth (0.056)

	(1)	(2)	(3)	(4)
	Reviewer's Rating	Reviewer's Rating	Reviewer's Rating	Reviewer's Rating
RD	-0.5241***	0.0021	-0.3852***	0.006
	(0.007)	(0.006)	(0.012)	(0.010)
diff	1.5478***	0.8761***	1.1021***	0.6204***
	(0.100)	(0.093)	(0.174)	(0.158)
intercept	0.6073***	0.5075***	0.5176***	0.4865***
	(0.015)	(0.014)	(0.004)	(0.004)
Controls	Yes	Yes	Yes	Yes
Observations	651775	714161	123679	141883

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) includes all users

(2) is a placebo test with imaginary rounding thresholds for (1)

(3) includes only elite users

(4) is a placebo test with imaginary rounding thresholds (3)

The results in table 6 still support the existence of attribution bias in the data since they are almost the same as the previous results. In column 1, I use the optimal bandwidth calculated based on the method in [47] and include all Yelp users within the bandwidth. The effect of *RD* is again negative and has a similar size as before. In column 3, I choose another optimal bandwidth with elite users only, and the effect of *RD* is negative and smaller than the effect in column 1. These results are consistent with the analyses in the previous section. In columns 2 and 4, I provide the placebo tests for columns 1 and 3. The effects of *RD* are almost zero and insignificant.

1.5 Discussion and Conclusion

1.5.1 Implication of Attribution Bias in Online Reputation Systems

From the regression results in the previous section, I provided empirical evidence of attribution bias on crowd-sourced review websites. To understand the implication and importance of attribution bias in this setting, I have linked my results to previous literature.

In [62], he showed that a half-star increase in ratings on Yelp can help restaurants' revenue by 5 to 9 percent. In [1], they provided another piece of evidence that Yelp ratings can help businesses and make the restaurants they studied sell out 49 percent more frequently. These results suggest that Yelp ratings can give strong incentives for business owners to improve the quality of their products and services. However, when I combine what they discovered with the findings in my paper, the incentives may be distorted by attribution bias. For instance, when a business's true quality is around one rounding threshold on Yelp, it may not always be good to improve its quality and cross the threshold. When the owner of the business only makes minimum effort to cross the threshold, attribution bias could cause other reviewers to review the business more harshly and make her effort in vain. In order to prevent this curse of attribution bias, the owner needs to make extra effort and make her business's quality significantly better than the threshold. This potential dynamic with attribution bias between crowd-sourced review websites and businesses was not credited in the literature before.

Attribution bias in the information aggregation process on consumer-reviewed websites is another important problem that has to be considered. It is usually believed that these websites benefit consumers by allowing them to collect useful information with low cost and decrease the information asymmetry in many markets. The information provided by the crowd-sourced forums is also believed to be reliable. Concerns about possible manipulations on these websites have been alleviated by [62] and [1]. In their papers, they utilized an econometric method proposed in [64] to detect manipulations of running variables in regression discontinuity designs and excluded the possibility of rating manipulations on Yelp. However, attribution bias is not a manipulation, and without considering its existence, the current online rating systems may not be credible. Although we may be able to argue that the averages of many consumers' ratings may still be robust and close to the true qualities, other users can still read biased reviews and receive inaccurate information as a result of attribution bias.

The above-mentioned implications of attribution bias in online reputation systems show us the importance of being aware of its existence and impacts. More studies about system dynamics and welfare effects with respect to attribution bias would help us know how this bias shapes economic activities and invent a better mechanism to gather consumers' opinions online.

1.5.2 Conlusion

This paper links to the developing literature of attribution bias in economics and provides the first empirical evidence of attribution bias from observational data. The results of this paper support the hypothesis that users have attribution bias when they review businesses. In addition, the paper points out potential problems of current online reputation systems. Future studies of attribution bias in different fields can be valuable since there are many other possible applications of attribution bias that can be explored. One potential direction that is closely related to this paper is utilizing text content in reviews to understand whether users also show attribution bias in other contexts. For instance, consumers may have different expectations when they go to a restaurant for special occasions. By studying the text reviews, we may be able to discover attribution bias in their dining experiences.

A broader direction worth pursuing is understanding the best way to provide quality measures. Presenting measures with rounding stars makes it easy for people to process the information. However, the findings in this paper suggest that attribution bias could affect people's perceptions and influence their opinions when we need their feedbacks. When businesses or governments try to simultaneously provide information and collect feedbacks by creating dynamic systems, how to strike a balance between user-friendly measures and precise information is an important policy question. Answering this question would create a more reliable way to take advantage of collective wisdom.

2.0 Hope Hurts: Attribution Bias in Yelp Reviews

This paper incorporates applied econometrics, causal machine learning and theories of reference-dependent preferences to test whether consuming in a restaurant on special occasions, such as one's birthday, anniversary, commencement, etc., would increase people's expectations and would make consumers rate their consumption experiences lower. Furthermore, our study is closely linked to the emerging literature of attribution bias in economics and psychology and provides a scenario where we can test two leading theories of attribution bias empirically. In our paper, we analyzed reviews from Yelp and combine the text analyses with regressions, matching techniques and causal machine learning. Through a series of models, we found evidence that consumers' ratings are lower when they went to the restaurants on special occasions. This result can be explained by one theory of attribution bias where people have higher expectations about restaurants on special occasions and then misattribute their disappointment to the quality of the restaurants. From the connection between our empirical analysis and theories of attribution bias, this paper provides another piece of evidence of how attribution bias influences people's perceptions and behaviors.

2.1 Introduction

It is long acknowledged that one's behavior and decisions are influenced by reference points (24). Prospect theory (48) indicates that an individual's overall utility is not solely determined by consumption experience, but also by one's reference level, like expectations. Expectations, as a commonly used reference point, is widely discussed in many economic studies in terms of its influence on people's behavior and decisions (51; 52; 69). The expectation (dis)confirmation theory (6; 71) is another similar theory existing in information systems and psychology. It predicts that higher prior expectations often come with higher likelihood of disconfirming beliefs. Both the prospect theory and the expectation (dis)confirmation theory predict that higher prior expectations are usually less likely to be met and often end up with lower satisfaction. In this study, we incorporated such theories with empirical data to explain a seemingly counter-intuitive phenomenon: hope hurts. Consuming in one's favorite restaurant on special occasions, such as birthdays, would more likely to result in dissatisfaction than to bring happiness due to higher expectations than usual days.

To start with, we constructed a reference-dependent utility model which incorporates psychological states. In the model, we allow states to interact with reference points, and this flexibility helps explain why special occasions may build up people's expectations about restaurants and make their dining experiences less enjoyable.

Empirically, we connected the user and business-level information with our theoretical framework and check whether special occasions affect Yelp users' behaviors. In order to uncover the causal relationship between special occasions and ratings, we employed applied econometrics and natural language processing techniques to glean insights from text reviews in combination with other variables.

Besides the average effect of special occasions, we are also interested in their heterogeneous impacts on different users. To account for this aspect, we applied the recent methods from casual machine learning, such as casual tree and causal forest to estimate the heterogeneous effects.

Our paper contributes to the emerging literature of attribution bias in behavioral economics. In previous papers, decision makers' attribution biases are either caused by differences between their expectations and true experiences (10; 46) or states when they make the decisions (36). In the scenario of special occasions, both scenarios can happen and influence Yelp users' behaviors in opposite directions. Thus, by using data from Yelp, we are able to study and understand what the possible theoretical and empirical results can be when these two causes of attribution bias interact with each other. Another contribution of this study is related to the ongoing discussion on how to create good crowd-sourced rating systems. Previous literature has shown drawbacks of current rating systems (16; 46) and our paper provides more insights on this topic. From another perspective, our paper suggests that ratings about consumption in different situations may provide distinct information. Finding good ways to help users digest information with considerations about review scenarios could potentially improve the usefulness of rating systems. Our results also bring practical insights to marketing strategies about how to launch successful marketing campaigns. With consumers' preferences in mind, businesses' owners could design promotions to avoid the negative effects of reference dependence by decreasing gaps between consumers' expectations and true experiences.

The rest of the paper proceeds as follows. In the next section, a detailed literature review is provided to connect our paper to existing research in behavioral economics, information system, and Psychology. In section 3, we propose a theoretical framework to discuss how different causes of attribution bias interact with each other. Our empirical strategy and data source are provided in section 4. In section 5, we present main estimation results and robustness checks for our analysis. Finally, discussions about implications of attribution bias and conclusions are included in the last section.

2.2 Related Literature

This paper connects several disciplines of research to answer questions in the intersection of behavioral economics, psychology and information systems. Thus, the literature review in this section gives a broad overview for relevant papers in the fields mentioned above.

In behavioral economics, scholars have discovered anomalies about how psychological factors affect decision makings of people. In [44] and [11], researchers found that weather influences people's housing and automobile purchasing decisions and the effects are explainable by psychological mechanisms but not classical utility theories. [81] and [68] provided empirical results and a theoretical framework about how investors' emotions influence their decisions and financial markets. Following these papers, our research studies the effect of potential psychological factors from special occasions, how they impact the information aggregation process on Yelp and the mechanisms behind them.

In discovering the potential mechanisms behind the influence of special occasions, a closely related literature in behavioral economics is about learning with misattributions. [50] developed a reference-dependent utility model which describes how reference points are formed and influence agents' behaviors when there is uncertainty. [30] extended the reference-dependent utility model and formalized attribution bias into their theoretical framework. In their model, the researchers described how decision makers can misattribute their gain-loss utility to their consumption experiences when the utilities they get are different from their expectations. [46] used the model of [30] and tested the existence of attribution bias on online reputation systems with Yelp's data. In [36], the authors proposed another model of attribution bias based on an extension of projection bias. From their model, states under which consumers make their consumption are another source of attribution bias. In our paper, we combined the two sources of attribution bias (30; 36) and developed a reference-dependent utility model with consideration of consumers' consumption states. This model not only helps us understand how two types of attribution bias interact with each other but also allows us to form testable hypotheses that we can check with review data from Yelp.

The Expectation Disconfirmation Theory (abbreviated as EDT below) is another strand that could account for attribution bias in online rating systems. EDT suggests that consumers would evaluate the difference between product performance and their expectations of the product (or anticipated performance of the product), and the "calculated gaps" influence their satisfactions. When the expectation of the product is higher, ceteris paribus, the likelihood of disconfirmation increases correspondingly and results in lower consumer satisfactions. In the fields of information systems (6; 9; 16; 39; 55; 65) and of marketing (2; 13; 20; 72; 73), this theory has already been widely used to describe the determinants of customer satisfactions, and the disconfirmation is often viewed as the mediator (20). In addition, expectation disconfirmations can happen in two types. When the product performance is better than the expectation, this is called "positive disconfirmation". On the contrary, if the performance is worse than the expectation, it is called "negative disconfirmation". In our study, we focus on cases where "negative disconfirmations" can potentially happen.

Theoretically, there are several theories in psychology that can provide mental mechanisms for expectation disconfirmations and attribution bias. A relevant one is called contrast effect (12; 63), which suggests that when negative disconfirmations happen, the perceived performance of products would be much worse, and vice versa. In our case, for example, when consumers' goal is to celebrate their birthdays in restaurants, their expectations are likely to be higher than usual due to the effect of special occasions. Following the higher expectations, the occurrence of negative disconfirmations becomes more likely and causes stronger dissatisfactions. In other words, during special occasions, the chance that Yelp users have negative dining experiences due to high expectations goes up. In this scenario, contrast effect triggers the dissatisfactions and causes users to give lower ratings for their consumption.

Other psychological mechanisms, such as assimilation-contrast effect (42) or generalized negativity (14), could also account for this cognitive bias. The underlying explanations are a bit different among these theories. However, they all provide similar predictions as our current study. Last but not least, how contrast effect magnifies the negative disconfirmation echoes the models in [30], and they all lead to the classical conclusion of reference-dependent preferences: losses loom (48).

The last stream of related literature is about disadvantages and improvements of word-of-mouth online reputation systems. Previous research has shown how multidimensional rating systems can lead users to write systematically different reviews compared to single-dimensional rating systems (16; 78). Others presented that the ubiquitous rounding feature in most online rating systems may lead to unexpected impacts on both reviewers and business (1; 46; 62). Our paper connects with these papers and shows that reviews from different scenarios should be considered separately because attribution bias plays an important role in them. A multidimensional rating system may potentially alleviate this problem since it can provide more background information about the reviews.

2.3 Theoretical Framework

In previous literature, potential biases due to difference between expectations and real experience are commonly explained with reference-dependent utilities (50) in behavioral economics or the expectation disconfirmatition theory (6, 9) in information systems.

In this section, we present a theoretical model extending those models and closely follow the settings from [30] and [36]. The model illustrates how Yelp users' consumption utilities are influenced by their states and their reference points when they make consumptions. The impacts on users' consumption utilities further influence their ratings on Yelp.

2.3.1 The Model

For a Yelp user i, her utility u_i, from a dining experience at time t is the following:

$$u_i(q|s_t, r_i) = \underbrace{v_i(q, s_t)}_{\text{Consumption utility at state } s_t} + \mu_i \underbrace{n_i(v_i(q, s_t)|r_i)}_{\text{Gain-loss utility}} \tag{8}$$

$$n_i(v_i(q, s_t)|r_i) = \begin{cases} f_i(|v_i(q, s_t) - r_i|) \text{ if } v_i(q, s_t) \geq r_i \\ \lambda_i f_i(|v_i(q, s_t) - r_i|) \text{ if } v_i(q, s_t) < r_i \end{cases} \tag{9}$$

Following 48, user i's utility is a combination of consumption utility and gain-loss utility in equation (1), and μ_i governs the weight she puts on her gain-loss utility. For her consumption utility, v_i, we assume it is a function of a restaurant's quality, q, and her state, s_t, at time t. s_t can be considered as a numerical measure of the user's average happiness level at time t. For instance, if user i visit a restaurant on a special occasion at time 0, and she revisit the restaurant again on a normal occasion at time 1, we will expect that $s_0 > s_1$. Moreover, better qualities and higher states are assumed to bring consumers higher utilities. To guarantee the property, we let v_i be a concave function with respect to q and s_i.

The gain-loss utility n_i is further explained in equation (2). $f_i(x)$ is a concave function increasing in x and $f_i(0) = 0$. The difference between the user's consumption utility, v_i, and her reference point, r_i, determines if there is a gain or loss in her dining experience. $\lambda_i < -1$ is assumed to capture loss aversion. When v_i is larger than r_i, there is a gain for user i, and the gain-loss utility is positive. On the other hand, if v_i is smaller than r_i, the gain-loss utility is negative.

With the model in this section, we can derive some interesting hypotheses with easy comparative statics. And these hypotheses allow us to test them empirically with Yelp's data.

2.3.2 Hypotheses from the Model

In this section, we use the model in the previous section to discuss possible effects of s_t and what implications of those effects are.

When we consider our model, if we only allow s_t to change and fix other variables, it is very natural to see that when s_t is larger, u_i is larger. For example, if user i goes to a restaurant twice with a same reference point, r_i, she will enjoy her dining experience more when she goes there on a special occasion.

To summarize this effect, we have our first hypothesis,

Hypothesis 1. *When $s_0 > s_1$, and everything else is fixed, we have*

$$u_i(q|s_0, r_i) > u_i(q|s_1, r_i)$$

.

Hypothesis 1 gives us an intuitive way to understand how s_t influences users' experience. However, it is also shown in the literature (15; 70) that r_i may not be independent of s_t, and higher s_t can lead to higher r_i. If we assume r_i is increasing in s_t, the effect of s_t will become ambiguous. To further explain possible effect under this new assumption, we have the next hypothesis.

Hypothesis 2. *When $s_0 > s_1$, $r_i(s_t)$ is increasing in s_t, and everything else is fixed, there are two possible outcomes.*

If $v_i(q, s_0) - v_i(q, s_1) > \mu_i(n_i(v_i(q, s_1)|r_i(s_1)) - n_i(v_i(q, s_0)|r_i(s_0)))$, then $u_i(q|s_0, r_i) > u_i(q|s_1, r_i)$.

If $v_i(q, s_0) - v_i(q, s_1) < \mu_i(n_i(v_i(q, s_1)|r_i(s_1)) - n_i(v_i(q, s_0)|r_i(s_0)))$, then $u_i(q|s_0, r_i) < u_i(q|s_1, r_i)$.

In hypothesis 2, we see that if the reference point r_i increases a lot because of the increase in s_t, it is possible for user i to have a lower utility in a higher state. For example, when user i has a very high expectation about celebrating her birthday in a restaurant, it is likely that she will be disappointed. The disappointment can lead to a experience worse than her previous visits.

The comparative statics give us a chance to test whether reference points change with psychological states and how do they interact with states empirically. These hypotheses lead to our identification strategies in the next section.

2.4 Data Source and Empirical Strategy

To test our theory empirically, we use data from Yelp Dataset Challenge round 13. In the data, we have access to all user reviews in 11 cities around the world, and the time window of the data is from October 2004 to November 2018. By using the data, we know users' previous rating history, rating dynamics of all businesses, text reviews from users and many other information which we can use as control variables in our regressions.

To address the question of interest and link the observational data with our theoretical framework, we first classified reviews into special and non-special occasions. To do so, we searched keywords related to special occasions, such as birthday, anniversary and commencement. When these keywords showed up in text reviews, we labeled the reviews as written on special occasions. This method may lead to some measurement errors. A detailed discussion is provided in next section. After identifying reviews on special occasions, we conducted our main analysis on *repeated reviews* on Yelp. *Repeated reviews* are reviews from users who had at least one consumption on non-special occasions and at least one on special occasions in the same restaurants. Because they are reviews from the same users who went to the same restaurants on different occasions, the unobserved heterogeneity of Yelp users and of restaurants are mainly controlled by the within subject comparison.

To begin with, we ran a series of ordinary least square (OLS) estimations to examine effects of special occasions. The main regression we study in our analysis is:

$$star_{ijt} = \alpha_0 + \alpha_1 SpecialOccasion_{ijt} + \alpha_2 R_{ijt} + \lambda X + \epsilon_{ijt} \tag{10}$$

$$SpecialOccasion_{ijt} = \begin{cases} 1 \text{ if the dining experience was on a special occasion} \\ 0 \text{ if not} \end{cases}$$

In equation 3, $star_{ijt}$ is the rating which user i gave on Yelp for her dining experience at restaurant j at time t. $SpecialOccasion_{ijt}$ is an indicator function equals to 1 when the dining at restaurant j at time t was on a special occasion. R_{ijt} is a potential reference point for user i at time t for restaurant j. In previous literature(46), it was shown that Yelp users use restaurants' average ratings as their reference points before their visits. In our analysis, we assume Yelp users gather information about restaurants on Yelp and form their expectations based on the information. For users who did not visit a restaurant before, their R_{ijt}s equal to the average ratings of restaurant j they saw on Yelp before first visits. For users who repeatedly visit a restaurant, the R_{ijt}s are their ratings for restaurant j from previous visits. X includes all other potential control variables, such as pricing ranges of the restaurants, average ratings of the restaurants and how many people like the reviews, etc.

The identification strategy behind equation (3) is based on hypothesis 1 and 2. From hypothesis 1, when the reference points for user i is independent of her state s_t, a higher s_t leads to a higher utility. If the assumptions of hypothesis 1 are true, we will expect the regression coefficient of $SpecialOccasion_{ijt}$ to be positive. On the other hand, we know from hypothesis 2 that if the q of a restaurant is fixed but r_i is not fixed and is a function of s_t, it is possible for user i's utility to go down when s_t goes up. If we find that the coefficient of $SpecialOccasion_{ijt}$ is negative after controlling other variables, we may reject hypothesis 1, and the most appealing explanation will be that r_i increases with s_t and the increase in the reference points leads to disappointments when users dine in the restaurants on their special occasions. Thus, equation (3) gives us a way to test our theory and hypotheses.

In addition to OLS, we used difference-in-differences estimators (5) to factor out the influence of order effects (of restaurant visits). For Yelp users who visit the same restaurants several times, there may be some trends for their ratings. Difference-in-differences (diff-in-diff) estimators allow us to control for those trends and make the comparison between special and non-special occasions more reliable. In our dif-in-dif analysis, we estimated the effect of special occasions on Yelp users who went to the same restaurants twice, and at least for one of their repeatedly-visited restaurants, the first visit was on non-special occasions and second visit was on special occasions. When focusing on these users, we were able to compare the restaurants they went on special occasions and non-special occasions and conclude what the effect of special occasions on their second visits is. For the dif-in-dif estimators, the regression we studied is the following:

$$star_{ijt} = \alpha_0 + \alpha_1 Treated_{ij} + \alpha_2 RpVisit_{ijt} + \delta RpVisit_{ijt} \times Treated_{ij} + \lambda X + \epsilon_{ijt} \qquad (11)$$

$$Treated_{ij} = \begin{cases} 1 \text{ if user } i \text{ has visited restaurant } j \text{ on a special occasion} \\ 0 \text{ if not} \end{cases}$$

$$RpVisit_{ijt} = \begin{cases} 1 \text{ if it is the second time user } i \text{ visited restaurant } j \\ 0 \text{ if it is the first time} \end{cases}$$

Here, $star_{ijt}$ and X are defined the same way as in equation (3). $Treated_{ij}$ is a variable to identify whether a Yelp user has been to the restaurant on a special occasion. $RpVisit_{ijt}$ shows if a dining experience is a user's first visit or not. The dif-in-dif estimator we are interested in is δ, the coefficient of $RpVisit_{ijt} \times Treated_{ij}$, which shows the effect of special occasion on users' second visits.

Besides above estimations, we also applied cross-classified multilevel models (56, 43) to account for the special structure of Yelp's data. It is worth noting that the relationship between users and restaurants are not necessarily hierarchical. For example, Yelp users may consume in different restaurants and hence, the structure of the data is not conventionally nested. A diagram explaining the data structure is provided in the appendix. To cope with this special data characteristic, we employed the cross-classified multilevel model in our study, which gives us better precision for our estimations.

Moreover, we are also interested in learning heterogeneous treatment effects of special occasions. To better understand how special occasions affected different subgroups in our data, we used causal tree (3) and causal forest (83) to estimate the heterogeneous effects.

To check the robustness of our analysis, we conducted similar estimation on a larger sample including all users who reviewed a restaurant repeatedly on non-special occasions. However, users who went to a restaurant on special occasion can be very different from those users who went there on non-special occasion. To avoid systematic differences for different kinds of users, we applied propensity score matching to control for various variables and make the two groups more comparable.

Another potential threat to our inference is reviewers' self-selection. It is probable that those who were extremely satisfied or unsatisfied were more likely to rate the restaurants and to post their reviews. Fortunately, we could attenuate this concern by using Yelp's elite feature. Yelp elites are those users who are the most active and professional. They tend to post comments for every restaurants they have visited. Using data focusing on Yelp elites not only minimizes self-selection threat, but also gives us an opportunity to examine whether elites are free from the reference-dependent bias. Furthermore, we can compare elite with non-elite users by estimate the heterogeneous treatment effects for both subgroups.

Lastly, some may question if there is a ceiling effect in the rating system. For example, consumers are very likely to choose better restaurants to consume on their special occasions. If these restaurants are already 5 star ones in their minds, even the experiences are beyond their expectations, there is no room for them to give ratings higher than 5 star. This could be a potential threat to our analysis since the effect of special occasion could just come from the ceiling effect. Therefore, we also run Tobit models to deal with this censored data situation.

2.5 Empirical Analysis and Results

In this section, we first show the results from our OLS estimations and dive into crossed-classified multilevel models to present more precise estimates when taking our data structure into consideration. After the first set of results, we show the estimates of heterogeneous treatment effects from our causal forest and causal tree estimations. More robustness checks are provided in the end of this section.

2.5.1 OLS, Diff-in-Diff Estimators and Multilevel Models

As we mentioned in the identification strategy, when we estimated equation (3), we focused only on *repeated reviews* from users who have visited same restaurants on both special and non-special occasions. In this way, we can have an apple to apple comparison since it is similar to a within subject quasi-experiment. The results is shown in table 7 with control variables.

The first column in table 7 shows the OLS result without control variables. The coefficient for $SpecialOccasion$ is negative and statistically significant at 0.01 level. To guarantee other factors influencing users' reviews are taking into consideration, we put in other control variables in the estimations of equation (3). One concern about column 1 is that people may act differently when it comes to reviewing restaurants in different star ranges. In column 2, we put in fixed effects for different average rating ranges. For instance, a restaurant with average rating at 3.5 star will be given a fixed effect for being in the 3.5 star range. The estimate still shows a negative effect of $SpecialOccasion$. These results are different from what we derived from hypothesis 1. In hypothesis 1, we expect $SpecialOccasion$ to have positive effect on reviewers' ratings. However, we see a negative effect of $SpecialOccasion$.

Another possible explanation of what we see in column 2 is that only users who were disappointed about their repeated visits would rate a restaurant more than once, and users

Table 7: OLS Results of Special Occasion Effect on *Repeated Reviews*

	(1) Reviewer's rating	(2) Reviewer's rating	(3) Reviewer's rating	(4) Reviewer's rating
$SpecialOccasion$	-0.2410***	-0.2404***	-0.3693***	-0.3374***
	(0.024)	(0.023)	(0.038)	(0.036)
$RpVisit$	NA	NA	-0.0759**	-0.1081***
	NA	NA	(0.038)	(0.028)
$SpecialOccasion \times RpVisit$	NA	NA	0.1957***	0.1673***
	NA	NA	(0.047)	(0.044)
Controls	No	Yes	Yes	Yes
Observations	12632	12632	12632	12632

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the OLS result without any controls

(2) shows the OLS result with fixed effects for different rating ranges

(3) shows the OLS result with column 2's fixed effect and controls for repeated visits

(4) adds potential reference points and more control variables

tend to celebrate their special occasions in restaurants they have visited before. If that is the case, the negative effects we see from *SpecialOccasion* are not about special occasions but about the order of users' visits. To eliminate this concern, we control the order of the visits of the Yelp users by adding a variable $RpVisit$. $RpVisit = 0$ when it is the first time a user visits a restaurant, and $RpVisit = 1$ when it's not the first time. By using $RpVisit$ and the interaction term of $RpVisit$ and *SpecialOccasion*, we see that $RpVisit$ does have a negative effect on people's ratings. However, the coefficient of *SpecialOccasion* is still negative and statistically significant. This result shows that the effect of *SpecialOccasion* is not just due to the order of users' visits.

To better fit our empirical analysis with the theory in the previous section, we also put in potential reference points of users in the estimation. To find good reference points, we assume Yelp users used the average rating of a restaurant as the reference points for their first visit, and they used the ratings they previously gave as the reference points when they visit the restaurant more than once. If reference points are fixed and does not change with users' states as the assumption of hypothesis 1, we should expect the effect from *SpecialOccasion* to disappear when we control for the potential reference points. However, we still see a negative effect of *SpecialOccasion* in the estimation result in column 4, which suggests that users' reference points actually changed with special occasions. The results in table 7 support our hypothesis 2 and shows that the effect size of increasing the reference points are larger than the effect size of being in a higher state.

For understanding potential heterogeneous effects on different groups of users, we conducted another analysis focusing only on the elite users of Yelp. Elite users are users who actively contribute on the platform and are dedicated in providing useful information to others. Those users tend to record their experiences as much as possible, and this feature gives us a chance to check what the impact of special occasions is on the most experienced users. In table 8, we show the results of elite users. For different control variable settings, the effect sizes of *SpecialOccasion* on ratings are smaller comparing to previous results with all users. However, the effect is still all negative and statistically significant. This shows that though elite users are more experienced, they are still not immune to attribution bias.

When we try to interpret the results from table 7 and 8, the effect of *SpecialOccasion* is

Table 8: OLS Results of Special Occasion Effect on *Repeated Elite Reviews*

	(1) Reviewer's rating	(2) Reviewer's rating	(3) Reviewer's rating	(4) Reviewer's rating
$SpecialOccasion$	-0.0826**	-0.0829***	-0.1341***	-0.1208***
	(0.033)	(0.031)	(0.048)	(0.046)
$RpVisit$	NA	NA	-0.0421	-0.1208***
	NA	NA	(0.04)	(0.028)
$SpecialOccasion \times RpVisit$	NA	NA	0.0880	0.0996*
	NA	NA	(0.061)	(0.058)
Controls	No	Yes	Yes	Yes
Observations	4612	4612	4612	4612

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the OLS result without any controls

(2) shows the OLS result with fixed effects for different rating ranges

(3) shows the OLS result with column 2's fixed effect and controls for repeated visits

(4) adds potential reference points and more control variables

ambiguous for Yelp users' non-first visits. Since we see the coefficients of $SpecialOccasion \times RpVisit$ are positive, when we compare special occasions with non-special ones for Yelp users' non-first visits, it is unclear if *SpecialOccasion* has significant negative effects on reviewers' ratings. To get more precise estimates of the impact of special occasions, we use difference-in-differences estimators to estimate the effect of special occasions on non-first visits. The estimations we do are based on equation (4), and the results are shown in table 9 and table 10. The coefficient of interest is $Treated \times RpVisit$, and we can see that when we include all users in table 9, the effects of special occasions on their second visits are negative and significant in all specifications. These results confirm that special occasions still cause Yelp users to give lower ratings even when they have previous experiences with the restaurants. For results in table 10, we see that the effect of special occasions is smaller for elite users, which is consistent with our previous OLS estimations. When we include all control variables, the influence of special occasions is still significant.

In our data with *repeated reviews*, users may repeatedly visit more than one restaurant. Thus, our review data is nested in both user level and restaurant level, and it has no clear hierarchy between user and restaurants. To account for this data structure and get more precise estimates, we used a cross-classified multilevel model (43) to conduct another series of estimations. In table 11, we still get similar results as before. All coefficients of *SpecialOccasion* are negative and significant at 0.01 level.

2.5.2 Heterogenous Treatment Effects with Causal Machine Learning

With the analyses above, we have tested the existence of attribution bias of high expectations. We also showed that *SpecialOccasion* could have differential effects on elite and non-elite Yelp users. To better understand the impacts of special occasions on different types of users, we employed a nonparametric causal forest algorithm (23; 83; 4) to examine the heterogeneous treatment effects. The main purpose of this analysis is to use causal forest to estimate conditional average treatment effect (CATE) of Yelp elites and non-elite users. CATE is defined as follows:

$$CATE : \tau(x) = E[Y_i(1) - Y_i(0)|elite_i = x],\ x \in \{1, 0\} \tag{12}$$

Table 9: Diff-in-Diff Results of Special Occasion

	(1) Reviewer's rating	(2) Reviewer's rating	(3) Reviewer's rating
$Treated$	0.2252***	0.1747***	0.1776***
	(0.025)	(0.024)	(0.023)
$RpVisit$	-0.1149***	-0.1149***	-0.1242***
	(0.013)	(0.012)	(0.011)
$Treated \times RpVisit$	-0.1114***	-0.1114***	-0.1756***
	(0.041)	(0.039)	(0.036)
Controls	No	Yes	Yes
Observations	37998	37998	37998

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the dif-in-dif result without any controls

(2) shows the dif-in-dif result with fixed effects for different rating ranges

(3) adds potential reference points and more control variables

Table 10: Diff-in-Diff Results of Special Occasion of Elite Users

	(1)	(2)	(3)
	Reviewer's rating	Reviewer's rating	Reviewer's rating
$Treated$	0.2172***	0.1752***	0.1927***
	(0.035)	(0.033)	(0.031)
$RpVisit$	-0.0754***	-0.0725***	-0.0866***
	(0.015)	(0.014)	(0.013)
$Treated \times RpVisit$	-0.0362	-0.0392	-0.1098**
	(0.054)	(0.051)	(0.047)
Controls	No	Yes	Yes
Observations	23300	23300	23300

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the dif-in-dif result without any controls

(2) shows the dif-in-dif result with fixed effects for different rating ranges

(3) adds potential reference points and more control variables

Table 11: Multilevel Model Results of Special Occasion Effect on *Repeated Reviews*

	(1)	(2)	(3)	(4)
	Reviewer's rating	Reviewer's rating	Reviewer's rating	Reviewer's rating
$SpecialOccasion$	-0.1679***	-0.1696***	-0.3058***	-0.3114***
	(0.019)	(0.018)	(0.036)	(0.035)
$RpVisit$	NA	NA	-0.1472***	-0.1438***
	NA	NA	(0.031)	(0.030)
$SpecialOccasion \times RpVisit$	NA	NA	0.2327***	0.2113***
	NA	NA	(0.050)	(0.047)
Controls	No	Yes	Yes	Yes
Observations	12632	12632	12632	12632

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the OLS result without any controls

(2) shows the OLS result with fixed effects for different rating ranges

(3) shows the OLS result with column 2's fixed effect and controls for repeated visits (4) adds potential reference points and more control variables

In equation (5), $Y_i(1) - Y_i(0)$ is the average treatment effect of $SpecialOccasion$. $elite_i = 1$ when the review is written by an elite user and $elite_i = 0$ if not.

Causal Forest is adopted here for a few reasons. Firstly, the assumption of linear interaction effect as the heterogeneous treatment effect in econometrics is very strong and often questioned (37). On the other hand, causal forest can automatically incorporate nonlinear functional form, such as higher-order terms or complex interaction effects, so the strong assumption of linearity is relaxed. In addition, casual forest is a delicate extension of the widely used algorithm of random forest (8); besides keeping the predictive capability of ensemble methods, causal forest further constructs asymptotic confidence intervals for the treatment effect, which acts as a great tool for the combination of causal inference and machine learning. Thirdly, causal forest as an improved version of machine learning technique provides out-of-bag prediction and attenuates the concern of over-fitting. Last but not the least, causal forest allows us to estimate heterogeneous treatment effects at group level, which fits our need perfectly.

Table 12: CATE from Causal Forest

	(1)	(2)	(3)
	ATE	CATE(Elite)	CATE(Non-Elite)
SpecialOccasion	-0.175	-0.064	-0.238
	(-0.214,-0.135)	(-0.123,-0.007)	(-0.291,-0.186)
Observations	12632	12632	12632

95% confidence intervals in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the average treatment effect (ATE)

(2) shows the conditional average treatment effect for elites

(3) shows the conditional average treatment effect for non-elites

Our results from causal forest estimation is presented in table 12 and figure 4. In table

12, the average treatment effect of special occasions is -0.175 and the 95% confidence interval is from -0.214 to -0.135. The size of the effect is similar to our OLS estimation and shows negative effect of $SpecialOccsion$ on reviewers' ratings. Moreover, the treatment heterogeneity is also detected as our previous analyses. Comparing with Yelp Elite, non-elite Yelp users suffer more from attribution bias (for the elites subpopulation, the confidence interval ranges from -0.123 to -0.007; for the non-elites subpopulation, the confidence interval is from -0.291 to -0.186). The magnitude of conditional average treatment effect (CATE) is stronger for these non-elite users (Figure 1), which indicates that they were more biased when they consumed on their special days. It is also worth noting that though the heterogeneous treatment effect is confirmed, the Yelp Elites are not free from this bias. In brief, even for the active and experienced users, attribution bias may still distort their judgment.

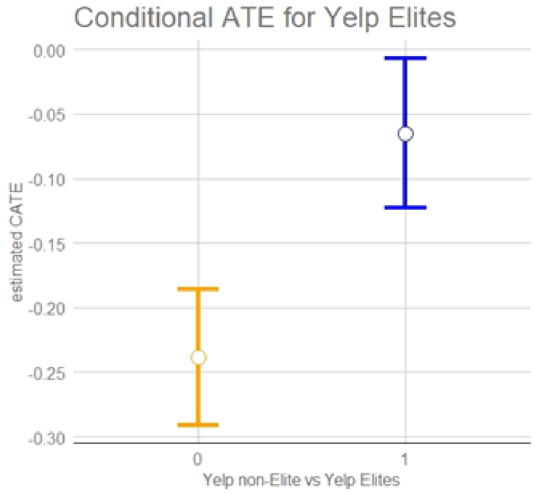


Figure 4: CATE for Yelp Elites

In addition, estimating the effect on Yelp elites has an additional benefit for our study: to attenuate the well-known extremity bias (38; 45; 58) in online rating systems. By comparing the rating distributions between elites and non-elites, we can observe that Yelp elite's proportions of choosing the extreme ratings (1-star and 5-star) are both lower than that of non-elites, ($\chi^2(4, N = 12632) = 623.48$, $p < 0.001$ (see Appendix)). Previous paper also shows that elite users do no try to revenge restaurants when there are potential gaps between restaurants' true quailities and their expectations in general (46). All of the evidence supports that elites suffer less from extremity bias. By estimating elite's CATE, it helps us examine the treatment effect when the extremity bias is less serious.

Another set of honest casual tree estimations of heterogeneous effects is provided in the appendix.

2.5.3 Robustness Checks

A potential threat to our OLS analyses is ceiling effect. Since Yelp users cannot acclaim restaurants with more than 5 stars, the ratings they can give are censored above. When users dine in restaurants with high average ratings on special occasions, even if they are very content with the restaurants, the most they can give is still 5 stars. This situation could make positive effects of special occasions undetectable and amplify the impact of negative reviews. This bias could also explain the negative coefficients of *SpecialOccasion* in our OLS estimations, invalidating our theory about attribution bias. To deal with this problem, we apply a Tobit model to attenuate the potential influence of ceiling effect. As the setting in [80], the observed dependent variable (Yelp user's rating) in our case is given by

$$y = y^* \text{ if } y^* < 5$$

$$y = 5 \text{ if } y^* \geq 5$$

Where y^* is the actual latent rating, and y is the observed rating. Because of the restriction on the rating scale, y cannot exceed 5, which is the highest possible score on Yelp. Thus, y^* is known exactly when it is less than 5 but unknown when it is greater than 5. To account for this data structure, we use censored regression models (Tobit Model) to analyze the data, and juxtapose the results of the whole sample and elite sub-sample with previous analyses. In table 13, we see the effects of *SpecialOccasion* are negative in all specifications, which supports our previous analyses. In our appendix, we also include estimation results for censored least absolute deviations estimators (75), which are also consistent with our previous results.

To see if our results can be generalized to a larger population, we included *repeated reviews* and reviews from users who have repeatedly visited the same restaurants on non-special occasions only. For example, if a user visited a restaurant twice on non-special

Table 13: Results of Special Occasion Effect on *Repeated Reviews* for Tobit Models

	(1) Reviewer's rating	(2) Reviewer's rating	(3) Reviewer's rating	(4) Reviewer's rating
$SpecialOccasion$	-0.4535***	-0.4612***	-0.7482***	-0.5952***
	(0.056)	(0.053)	(0.089)	(0.073)
Controls	No	Yes	Yes	Yes
Observations	12632	12632	12632	12632

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the Tobit result without any controls

(2) shows the Tobit result with fixed effects for different rating ranges

(3) shows the Tobit result with column 2's fixed effect and controls for repeated visits

(4) adds potential reference points and more control variables

occasions only, the reviews she wrote would be included in this larger sample. Since we do not necessarily have a within subject comparison in this sample, to account for potential bias in other variables, we used propensity score matching to estimate the causal effects of special occasions and the result is in table 14. From the estimate from the matched sample, the effect of *SpecialOccasion* is still negative and significant, which indicates that attribution bias also shows up in this larger sample.

Table 14: Matching Results

	Reviewer's Rating
SpecialOccasion	-0.0650***
	(0.0236)
intercept	3.8483***
	(0.0167)
Observations	133333

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

propensity score matching (knn)

In our previous analyses, we use keyword searching to identify whether Yelp users' reviews are about their dining experiences on special occasions. However, this method could cause potential measurement errors. For example, if a reviewer wrote "there were a family celebrating a birthday", the review will be classified as related to special occasions. We alleviated this concern by randomly selecting thousands of reviews and manually classifying them into special-occasion reviews or non-special ones. The results of this manually classified data are shown in table 15. From the table, we see that coefficients in the first two columns are with the same signs as our previous OLS analyses. Thus, even with some measurement errors, there seems to be no systematic bias because of the errors. A byproduct of this manual classification is that we are able to identify whether the reviewers celebrated their own special occasions or participated in celebrations of other people's special occasions. We took advantages of this finer classification and explored the differences between different types of special occasions. In column 3 and 4 of table 15, we present the effect of celebrating one's own special occasions and celebrations for others. The results show that when celebrating for others, the effects of special occasions are negative, and the sizes are significantly larger than celebrating one's own special occasions. From our observations when doing manual classification, these results are also caused by reference-dependent type of attribution bias. When arranging or recommending restaurants for others' special occasions, Yelp users' apply higher standards for food qualities and services. Such expectations make them more easily disappointed, and lead to lower ratings after their dining experiences.

Table 15: OLS Results of Special Occasion Effect on Manually Classified Data

	(1) Reviewer's rating	(2) Reviewer's rating	(3) Reviewer's rating	(4) Reviewer's rating
SpecialOccasion	-0.3976***	-0.3641***	NA	NA
	(0.059)	(0.051)	NA	NA
One's own special occasion	NA	NA	-0.1873***	-0.2059***
	NA	NA	(0.071)	(0.059)
Others' special occasion	NA	NA	-0.5773***	-0.5199***
	NA	NA	(0.077)	(0.069)
Controls	No	Yes	No	Yes
Observations	1847	1847	1847	1847

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the OLS result without any controls

(2) shows the OLS result with control variables

(3) shows the OLS result without any controls

(4) shows the OLS result with control variables

2.6 Discussion and Conclusion

2.6.1 Attribution Bias and Online Review Systems

As we briefly discussed in the literature review section, attribution bias caused by special occasions can be classified into a broader category of problems of online rating systems. In [16], the researchers provide evidence showing that multidimensional rating systems is more informative than systems only presenting average ratings. The cost for users to find relevant information on multidimensional systems is much lower since it automatically provides other reviewers' opinions from many aspects. Hence, the authors suggest that multidimensional rating systems can improve the information aggregation process of the word-of-mouth platforms. When considering their proposal together with our findings, multidimensional systems could potentially attenuate attribution bias as well. If more background information about reviews is provided in an easily accessible way, users will understand more about the scenarios when reviewers dined in the restaurants. Taking those extra information from multidimensional systems into account, the gap between users' expectations and experiences could be reduced, and the impact of attribution bias would be much smaller.

2.6.2 Practical Implication in Marketing

In this research, besides the estimation of average treatment effect of special occasions, we specifically examine the heterogeneity of treatment effect (HTE) for Yelp Elites and non-elites users. By applying the causal forest algorithm, the treatment heterogeneity is confirmed. The empirical results reveal that when encountering an expectation-experience gap (especially negative disconfirmation), the non-elite user's disconfirmation effect is much larger, which is denoted by the greater magnitude of the negative coefficient. In addition, it is worth noting that, even for the Yelp Elite users, the subpopulation who are active, experienced and professional for rating restaurants, they still suffer from the cognitive bias due to a higher expectation. In our view, this confirmed treatment heterogeneity for the Yelp elites/non-elites is not only for satisfying academic curiosity, but also offers insights for reflecting campaigns and opportunities for customizing organizations' marketing strategy.

Launching marketing campaigns or advertisements could be a double-edged sword. Undoubtedly, an effective campaign or an eye-catching advertisement can help attract more consumers and gain more profits. However, from the perspective of attribution bias or expectation disconfirmation, for those who are allured, their expectations are likely to be much higher than usual. This phenomenon creates a paradoxical circumstance: consumers who are attracted are now harder to feel satisfied. The tension between attracting more customers and increasing the likelihood of dissatisfaction seems to be a dilemma.

Nevertheless, considering the treatment heterogeneity of Yelp elites/non-elites in this study, this tension could be relieved or even be utilized. For example, it is a conventional practice for restaurants to release discounts or relevant information to attract consumers who want to celebrate their own birthdays. Based on the results we obtained in this study, it is already known that Yelp elites' (negative) disconfirmation effect is not as strong as those who are non-elites. Therefore, the businesses could make use of the different levels of expectation between experienced diners and new customers by offering discount information for the former before consumption (as usual discount activities), but offering the same information for the latter after consumption (serving as a surprise, beyond their expectations). By doing so, the businesses can simultaneously maximize the attraction of campaigns or advertisements while minimizing the concern of consumers' negative disconfirmation.

Furthermore, the tools we applied open the door to personalized marketing. Causal machine learning algorithms like generalized random forests (GRFs) (4), causal tree (3) and causal forest (83) allows us perform a sufficiently fine-grained level of analysis, estimating the user-level treatment effects. With sufficient user data, combining with predictive techniques and behavioral science knowledge, we could have a much deeper understanding of individual consumer preference and behaviors, which is pivotal for companies' marketing strategies. Needless to say, the combination of causal inference and machine learning models shows a promising way for organizations who are eager to make informed decisions from data.

2.6.3 Conclusion

In our paper, we show that attribution bias can be detected in online rating systems. The theoretical framework and the empirical results give us a new perspective on how reference dependence and state dependence create attribution bias and how they interact when it comes to reviewing businesses. Our research extends the previous studies about attribution bias in behavioral economics and further distinguishes different possible explanations. This paper also contributes to the discussion on how to form better rating platforms. From our analysis, we show that information on rating websites is highly related to the situations where users make their consumption. How to include useful information about the review scenarios while keeping overall content easy to digest is a question which needs further exploration. Furthermore, we connect our findings with practical application in marketing to provide potential strategies for businesses in the future.

Still, there remain some limitations unsolved in our paper. According to past studies (2; 12), there are several psychological models that could explain the effect of disconfirmed expectancy, like assimilation effect (29), contrast effect (42), generalized negativity (14), and assimilation-contrast effect (42). The latter three give the same prediction as our current results for the disconfirmation phenomenon.

Assimilation effect predicts that when users find gaps between their expectations and experiences, they have the tendency to solve the psychological discomforts because of the difference. This contradicts with our findings. The contrast effect (42) indicates that the expectation-experience gap would be magnified when disconfirmation happens. So when the users' experiences are inferior to what they expected, the negative disconfirmation makes them more dissatisfied. Secondly, the theory of generalized negativity (14) predicts that any sorts of disconfirmation, regardless of the fact that consumers confront the positive or the negative disconfirmation, they would feel unpleasant or unsatisfying due to the discrepancy. Lastly, the assimilation-contrast effect (42) offers a more complicated mental mechanism. It combines the assimilation effect and the contrast effect. It predicts that when the expectation-experience gap is small, the assimilation effect dominates, and when the gap is large, the contrast effect rules. In our case, it seems that the mixed prediction may undermine the likelihood of detecting the treatment effect of special occasions, because some Yelp users may perceive the expectation-experience gap small, and some may consider it large. However, we think the effect of special occasion makes these users hold much higher expectation ("Today is a special day!") for the coming consumption, and it is reasonable to anticipate that people will choose better restaurants for their special days, which also raises their expectations. All of these will increase the likelihood of negative disconfirmation, so we gauge that the large expectation-experience gap would be more prevalent in our case.

To sum up, with the results we obtained in the current study, it is not easy to distinguish between the different psychological models that could all explain the phenomenon we observe. Due to the limit of our expertise, we only provide the theoretical framework and analysis of one possible explanation - reference dependence and attribution bias. We welcome suggestions and future discussions on the other possibilities.

3.0 From Econometrics to Machine Learning: Application of Recurrent Neural Networks on Yield Curve Forecasting

Financial derivatives and interest rates correlate strongly with United States government bonds. Among many characteristics of government bonds, the term structure or the so-called yield curve is one of the main targets that investors always attempt to forecast. In this paper, I construct a model with recurrent neural networks (RNN) and focus on the point forecasting of the yield curve to explore the possibility of having a better forecast for the term structure. In addition, the similarities between RNN and the state-space models allow me to show that the newly proposed neural-network method is closely linked with previous financial econometric forecasting literature and can be considered as a generalization of the dynamic Nelson-Siegel method (Diebold and Li, 2006). While allowing similar interpretation as previous econometric methods, the neural network model in this paper shows better forecasting accuracy.

3.1 Introduction

The advance of machine learning in the past decades has created many possibilities in solving forecasting problems. Among machine learning approaches, neural networks have a particular advantage in capturing nonlinear relationships in data and are widely used in forecasting. When it comes to forecast time series data, one type of neural network, the recurrent neural network, has gotten increasing attention because of its dynamic flexibility in time domains. An ability to learn the importance of different data features in sequences allows a recurrent neural network to refine its forecasting results sequentially based on new input data and previous history. This characteristic is similar to many dynamic time series forecasting models in statistics and econometrics. However, in most traditional models, certain functional forms need to be assumed in advance. A recurrent neural network, on the other hand, is an assumption-free method and the model simply learns functional forms from data in its training stage. The similarities and differences between recurrent neural networks and traditional time series econometrics make them comparable and can help us understand what would be meaningful applications of recurrent neural networks in the financial forecasting.

Although scholars have created a huge body of literature on application of recurrent neural networks in previous decades, most of the applications are on machine translation and language parsing. It was only until recently that researchers have started to rigorously discuss how we can apply recurrent neural network to financial time series forecasting. This research is proposed to extend the boundary of current literature. Constructing forecasts for financial data is known to be one of the most difficult tasks among all forecasting challenges. The unpredictability of financial time series makes the classic random walk model remain a benchmark for point forecasting. Despite the random walk simply uses the current value as a forecast for the future, other time series models can hardly surpass it. This difficulty makes most traditional financial forecasting models undesirable since their results are not even better than the random walk.

The above-mentioned advantages of recurrent neural network and problems of time series models have inspired me to start this study. The paper aims to provide suitable neural network settings for financial time series forecasting. To make this research applicable and testable, I chose yield curves of United States government bonds as targets for proposed forecasting models and will test their performance based on forecasting errors of the government bonds' interest rates. This decision makes the study closely connected to practical needs of central banks and financial industries since global economies and financial derivatives correlate strongly with United States government bonds. Having a good point forecasting of yield curves is important for bond portfolio management, and a reliable density forecasting of interest rate is crucial for asset pricing and risk management. In this research, I focus on the point forecasting of the yield curve and compare different methods to explore the possibility of having a better forecast for the term structure.

3.2 Literature Review

This study is designed to answer questions in the intersection of machine learning, financial economics, and time series econometrics. Therefore the following review will cover literature in these three fields.

Machine learning with recurrent neural network can be dated back to 1980s. In their seminal paper, [77] introduce a recurrent neural network for the first time as a self-organizing learning mechanism in mimicking brain behaviors. [40] improved the work of Rumelhart et al. and proposed a long short term memory (LSTM) recurrent neural network. Their new model fixed the difficulty of gradient exploding in the original network and made LSTM more applicable to learning with long time steps. Since then, LSTM has been used to solve problems in different fields. For example, in improving speech recognition with artificial intelligence, [34] showed how LSTM can help decrease the recognition error rate. As another example, [18] modified LSTM and proposed an encoder-decoder mechanism to improve statistical machine translation.

In previous finance literature, the yield curve forecasting has been done using no-arbitrage models, dynamic stochastic general equilibrium (DSGE) models or econometric approaches without equilibrium and no-arbitrage conditions. Starting from [82], a series of papers have used either diffusion processes or discrete Markov affine models with no-arbitrage conditions to describe the term structure movements. [27], [33] fall into this category and follow the no-arbitrage tradition with different focuses. Approaching the problem from a different perspective, [22] laid the microeconomic foundation of the movement of yield curves from investors' preferences and constructed the DSGE outcome. Although the above-mentioned papers are theoretically appealing for explaining market behaviors, they were not designed for forecasting. According to [27], they all forecast poorly.

On the other hand, the econometric approaches without no-arbitrage and DSGE constraints perform better for the point forecasting task. In the literature along this line, [25] has been one of the most influential pioneers. In their paper, they use a latent dynamic factor model to fit the cross-sectional yield curve and dynamically update the factors to conduct forecasting. Compared to other competing approaches, this method is not only simple and elegant but it also improves the forecasting accuracy significantly. Following the breakthrough of Diebold and Li, [19] connected their econometric model to the no-abitrage constraints and showed that the model actually satisfies many properties in the finance theory of term structures. The research of Christensen et al. validated the application of Diebold and Li's model and bridged the gap between pure statistical approaches and finance theories on yield curves forecasting.

Recently, many scholars have started to work on financial forecasting using machine learning techniques. Forecasting economic outcomes with neural networks, [76] built a novel model with an attention mechanism on the foundation of the encoder-decoder structures and used the model in forecasting the NASDAQ stock index. [21] borrowed tools from natural language processing and applied the encoder-decoder architecture to forecast the United States unemployment rate and demonstrated that their models made better forecasts than financial experts. Connecting to financial theories more closely, [53] used a feedforward neural network and imposed no-arbitrage conditions to provide a general setting for conducting a yield curve forecasting with neural network. [7] proposed a neural network in forecasting excess bond return and presented statistical evidence in favor of their model. Although the emerging trend of combing financial forecasting and machine learning has led financial econometrics to a new direction, forecasting yield curve using recurrent neural networks still seems to be a missing part in the existing literature.

In this paper, I proposed a new application of recurrent neural network in yield curve forecasting. In particular, I compare several forecasting methods including random walk and dynamic Nelson Siegel model from [25] with my newly proposed recurrent neural networks in this paper.

The rest of the paper proceeds as follows. In the next section, I provide an introduction to the neural networks I used in this paper. The model setting of [25] and other methodologies are also presented in section 3. A detailed description of the data and empirical analyses are in section 4. In the end, I offer final discussions and conclude the paper.

3.3 Dynamic Nelson-Siegel Model and Neural Networks

3.3.1 Dynamic Nelson-Siegel Model

The Dynamic Nelson-Siegel (DNS) method was proposed by [25]. The model extends the work of [67] and integrates the dynamic features into it. This allows the model to use past data and forecast forward. The setting of the model is the following:

Assuming a cross-sectional model for yield curve at time t

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t}(\frac{1-e^{-\lambda\tau}}{\lambda\tau}) + \beta_{3,t}(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}) \tag{13}$$

y_t is the yield rate. τ represents the length of the interest rate we consider and can be 1 month, 2 months, 1 year, 30 years...etc. The three βs can be interpreted as three latent dynamic factors and $\{1,\ \frac{1-e^{-\lambda\tau}}{\lambda\tau},\ \frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\}$ are the loadings on those factors. λ is a macroeconomic parameter and can be calibrated with other macroeconomic models. In this paper, I follow the setting of [25] and let $\lambda = 0.0609$

The first stage of forecasting is a simple ordinary least square estimation. The model fits the historical data and estimates $\{\beta_{1,t},\ \beta_{2,t},\ \beta_{3,t}\}$ for each time point t. Once the estimated $\{\hat{\beta}_{1,t},\ \hat{\beta}_{2,t},\ \hat{\beta}_{3,t}\}$ are generated from the model, we proceed to the second stage.

In the second stage, an first-order autoregressive model is used to forecast the βs. Again, we conduct another ordinary least square estimation to get estimates $\hat{c_i},\ \hat{\gamma_i}$ of $c_i,\ \gamma_i$ from

$$\hat{\beta}_{i,t} = c_i + \gamma_i\hat{\beta}_{i,t-h}$$

and forecast $\hat{\beta}_{i,t+h}$ using

$$\hat{\beta}_{i,t+h} = \hat{c_i} + \hat{\gamma_i}\hat{\beta}_{i,t}$$

After these steps, the forecast $\hat{y}_{t+h}(\tau)$ is formed by plugging in $\{\hat{\beta}_{1,t+h},\ \hat{\beta}_{2,t+h},\ \hat{\beta}_{3,t+h}\}$. And we get

$$\hat{y}_{t+h}(\tau) = \hat{\beta}_{1,t+h} + \hat{\beta}_{2,t+h}(\frac{1-e^{-\lambda\tau}}{\lambda\tau}) + \hat{\beta}_{3,t+h}(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}) \tag{14}$$

From the above procedures, we can see that the model from [25] is both flexible and parsimonious. There are several advantages of the model. The number of parameters is relatively small in this setting and it makes estimations less computationally intense. The model can also fit the historical yield curve very well with the flexibility of exponential loadings. To forecast forward, the setup is just another first order autoregressive regression and can be easily conducted. The dynamic Nelson-Siegel model is also more interpretable compared to ARIMA model. Since the latent factors can be interpreted as a constant term, a slope term, and a curvature term, the model also tells us about the relative importance of short-term and long-term interest rates and can help us understand the current situation of government bonds markets. The forecasting results of dynamic Nelson-Siegel model is shown in section 4.

3.3.2 Recurrent Neural Networks

A recurrent neural network (RNN) is an assumption-free method, and the model simply learns functional forms from data in its training stage. The similarities and differences between RNN and time series econometrics make them comparable and can help us understand what would be meaningful applications of a recurrent neural network in the field of forecasting. Among the recurrent neural networks, Long Short-Term Memory (LSTM) Neural Network, a special type of them is widely used in solving the task of machine translation and language parsing. Recently, LSTM also becomes one of the most popular candidates when scholars attempt to use machine learning in time series forecasting because of its ability to capture useful information and forget unnecessary noises from the historical data. In this paper, the possibility of applying LSTM to yield curve forecasting is explored.

3.3.2.1 Simple RNN

Before introducing LSTM, a common structure of a recurrent neural network is presented first. A recurrent neural network takes an input sequence $\{x_t\}$ and feeds elements of $\{x_t\}$ into the network one at a time. Each time a x_t goes into the network, x_t is combined with previous hidden state s_{t-1} to construct a new hidden state s_t. With the new hidden state, the network generates an output o_t of time t with s_t.

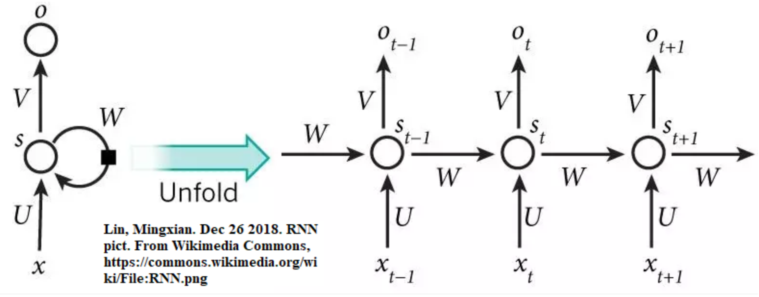
The details of the computation in a RNN are the following:

$$
\begin{aligned}
& s_t \in \mathbb{R}^I,\ x_t \in \mathbb{R}^P,\ o_t \in \mathbb{R}^K \\
& s_{t_i} = f(\sum_{j=1}^{P} U_{ji} x_{t_j} + \sum_{j=1}^{I} W_{ji} s_{t-1_j} + b_i), \quad i = 1, 2 \dots I \\
& o_{t_k} = g(\sum_{j=1}^{I} V_{jk} s_{t_j} + c_k), \quad k = 1, 2 \dots K
\end{aligned}
$$

A more compact way to write the equations above is:

$$
\begin{aligned}
s_t &= f(Ux_t + Ws_{t-1} + b) \\
o_t &= g(Vs_t + c)
\end{aligned}
$$

The hidden state s_t is a real vector in a I-dimensional space. Input x_t is a P-dimensional vectors and output o_t is a K-dimensional vector. When input x_t and previous hidden state s_{t-1} go into the RNN, they interact with coefficient matrices U and W. Each x_{t_j} in vector x_t is weighted by U_{ji}, and s_{t-1_j} is weighted by W_{ji}. In the end, those terms are summed up with a constant term b_i and go through a function f to create the ith element of a new hidden state as s_{t_i}. At time t, each element s_{t_j} in the hidden state vector s_t is weighted by V_{jk} and combined with a constant term c_k. The summation of weighted s_{t_j} and c_k is then transformed by a function g to create the output o_t. The above description of RNN is visualized as a graph by [59] as figure 5.



Lin, Mingxian. Dec 26 2018. RNN pict. From Wikimedia Commons, https://commons.wikimedia.org/wiki/File:RNN.png

Figure 5: Recurrent Neural Network

In my RNN, f and g are smooth functions with first and second derivatives. The elements in coefficient matrices U, W and V are the parameters to be estimated. When it comes to estimate the parameters, a loss function needs to be set up. In my application, I choose mean square errors of my forecasts as the loss function. In the training stage, parameters can be estimated with stochastic gradient descent mechanism since the derivatives of f and g exist.

3.3.2.2 LSTM

A LSTM neural network is built on simple RNN and was originally designed to solve the gradient vanishing and explosion problem of simple RNN. A LSTM makes further assumptions on the functional forms of f in previous section. It consists of five parts: the input gate, output gate, forget gate, memory cell and visible state. Here I follow the notation from [61] and show their exact forms as the following:

$$\begin{aligned}
\mathfrak{f}_t &= \sigma(W_{\mathfrak{f}} x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
\mathfrak{o}_t &= \sigma(W_{\mathfrak{o}} x_t + U_{\mathfrak{o}} h_{t-1} + b_{\mathfrak{o}}) \\
\tilde{c_t} &= tanh(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \otimes c_{t-1} + i_t \otimes \tilde{c_t} \\
h_t &= \mathfrak{o}_t \otimes tanh(c_t)
\end{aligned}$$

Here, $W_{\mathfrak{f}}$, W_i, $W_{\mathfrak{o}}$, $W_c \in \mathbb{R}^{h \times d}$ and U_f, U_i, $U_{\mathfrak{o}}$, $U_c \in \mathbb{R}^{h \times h}$ are the coefficient matrices. b_f, b_i, $b_{\mathfrak{o}}$, $b_c \in \mathbb{R}^h$ are constant vectors. Moreover,

$x_t \in R^d$: input vector to the LSTM unit

$\mathfrak{f}_t \in R^h$: forget gate's activation vector

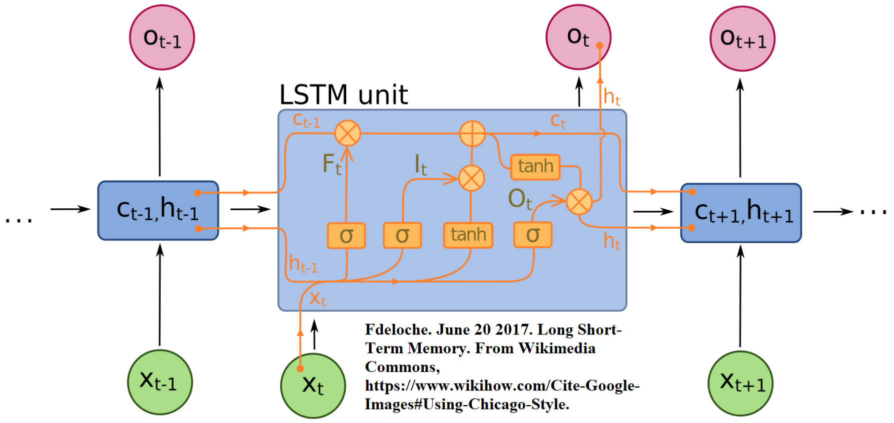
$i_t \in R^h$: input/update gate's activation vector

$\mathfrak{o}_t \in R^h$: output gate's activation vector

$h_t \in R^h$: hidden state vector, also known as output vector of the LSTM unit. This is the s_t in previous section

$\tilde{c}_t \in R^h$: cell input activation vector

$c_t \in R^h$: cell state vector



Fdeloche. June 20 2017. Long Short-Term Memory. From Wikimedia Commons, https://www.wikihow.com/Cite-Google-Images#Using-Chicago-Style.

Figure 6: LSTM

Although these formulas seem very complicated, the above computations can be visualized into simpler pieces in figure 6 from [28]. In figure 6, we can consider the big box as the functions f and g in the previous section and the inputs of the box are only c_{t-1}, h_{t-1} and x_t. The most important innovation of LSTM is the forget gate f_t. By choosing appropriate parameters in W_f and U_f, the problem of having vanishing or exploding gradients is greatly alleviated. More details of gradient problems are in [32] and [74]. To sum up, although LSTM uses an innovative and complicated way to solve the gradient problems, the structure is still the same as the simple RNN in previous section. In each time step, new input x_t is combined with previous hidden state h_{t-1} to create new hidden state h_t and output o_t through some potential nonlinear functions f and g.

3.3.2.3 Dual-Stage Attention-Based Encoder and Decoder

Before going into the actual model of this paper, another important structure, encoder and decoder, is introduced in this subsection.

Since there is no restriction on functional forms when training a RNN, we cannot guarantee the convergence of the model. Even if the model can converge, it may takes too long to achieve the goal. Thus, if we can utilize some prior knowledge about yield curve forecasting from previous literature, it could potentially be very helpful.

In [25], they showed the hidden states β_t are highly correlated with a constant level, slope and curvature of the yield curve. Moreover, when doing a principal component analysis on yield curve $y_t(\tau)$, the first three components are basically the three elements in β_t. By using this information, I was inspired to use an encoder-decoder architecture in my model.

An encoder-decoder architecture takes in data and transforms the data into lower dimensional features by using a encoder. With the lower dimensional features, we can decode the data by going through another decoder. This structure aims to find a lower dimensional representation of the data without losing important features of the data. In that sense, it is similar to principal component analysis and can be considered as a nonlinear PCA. The encoder-decoder structure is visualized as figure 7 from [66].

The encoder-decoder neural network is also used by many other researchers in finance and macroeconomics recently. In [35], they proposed usage of an encoder to incorporate information from other variables with historical returns of assets to improve their pricing accuracy. In [21], an encoder-decoder architecture is used to improve the forecast of United States unemployment rate.

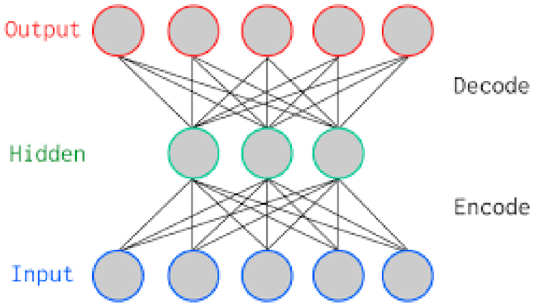


Figure 7: Encoder and Decoder

In my model, I use an encoder-decoder structure proposed by [76], which add two attention mechanisms into a two layered LSTM.

More details about this attention-based encoder and decoder are provided in the appendix.

3.3.2.4 Structure of the Final Model

In this section, I integrate all the tools in previous discussions and describe the model structure of this paper.

The first part of my model is a first-stage estimation of βs as the case in [25]. All yield rates y_t are compared with the output $\hat{y}_t$ by calculating mean square errors. Mean square errors are set up as the loss function and the estimation is conducted to minimize the mean square errors. The goal for this step is to encode yield rate into lower dimensions as hidden states.

The second part of the model is a dual-stage encoder and decoder similar to [76] that takes in hidden states (βs) from previous steps and generates outputs through the neural network as equation (3). The outputs of the dual-stage LSTM can be considered as forecasts of future hidden states. The forecast of future hidden states is then fed into equation (2) to create final forecasts. A loss function using mean square errors between the true yield rates and the forecasts is applied in the training stage, and stochastic gradient descent algorithm is conducted to minimize the mean square errors.

The structure of my model is visualized in figure 8, and figure 9. The training and testing process is summarized in algorithm 1.

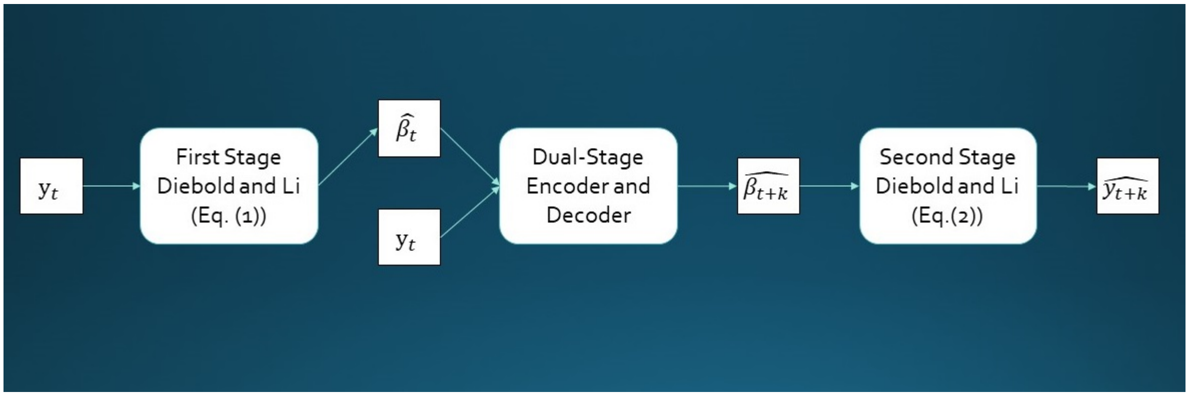

y_t
First Stage Diebold and Li (Eq. (1))
$\widehat{\beta}_t$
y_t
Dual-Stage Encoder and Decoder
$\widehat{\beta_{t+k}}$
Second Stage Diebold and Li (Eq.(2))
$\widehat{y_{t+k}}$

Figure 8: Final Model for k-period Ahead Forecasting

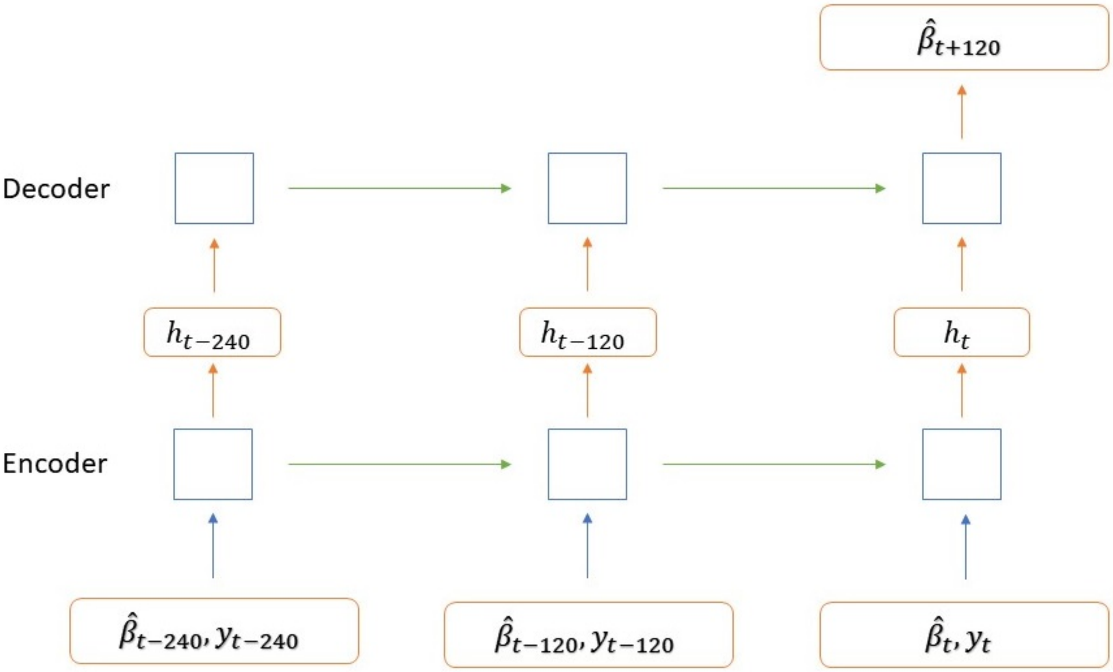


Figure 9: An Example of Dual-Stage Encoder and Decoder

Algorithm 1: Forecasting k-period ahead yield rate

```
Initialize all parameters;
Estimate equation (1) to get the hidden states ($\beta$s)
while forecast for the training set does not converge do
    plug in $\beta_{t-k}$ and $y_{t-k}$ into the dual-stage encoder and decoder to get $\hat{\beta}_t$;
    plug in $\hat{\beta}_t$ into equation (2) to get $\hat{y}_t$;
    calculate loss (mean-squared errors) from $y_t$ and $\hat{y}_t$;
    minimize loss function by doing stochastic gradient descent;
    update parameters of the dual-stage encoder and decoder;
    if meet convergence criterion then
        break the loop;
    end
end
fix parameters of dual-stage encoder and decoder;
conduct out-of-sample forecasting for $y_{t+k}$ based on the trained model
```

3.4 Empirical Analysis

3.4.1 Data Source

The main source of the data in this paper is constant maturity zero coupon Treasury yields provided by [60]. In their paper, a non-parametric kernel-smoothing method with a novel adaptive bandwidth was used to construct more reliable yield rates. The dataset is provided on their website [1] and the yield curve data is daily from June 14, 1961 to December 31, 2019. This novel dataset consists of yield rates with 360 different maturities in 14000 transaction dates. The size of this data allows me to apply complex neural network for my forecasting tasks with more confidence since training neural network requires estimating many more parameters than traditional time series methods and needs a larger sample size.

[1] https://sites.google.com/view/jingcynthiawu/yield-data

3.4.2 Forecasting Results

In this section, I show out-of-sample forecasts that are 120-transaction-day-ahead (6-month ahead) or 240-transaction-day-ahead (1-year ahead) for yield rates with maturities of 3 months, 12 months, 36 months, 60 months and 120 months from 1994/01 to 2000/12 and from 2015/01 to 2019/12. The gap between two neighboring forecasts is 20 transaction days. The forecast-accuracy measure used here is root-mean-square error (RMSE) and the results are shown in table 16 to 21. Evolution of cumulative RMSEs and forecast errors for each period are shown in the appendix.

In the previous study (25), the forecast is conducted with monthly data from 1994/01 to 2000/12. Thus, the first set of forecasts in this section serves as a robustness check of Diebold and Li's results and provides comparison between my proposed model and the dynamic Nelson Siegel (DNS) model. The second set of forecasts works as another evaluation of my model with a more recent interest rate data.

From now on, the proposed model in this paper will be called LSTM for convenience. The first set of results for 1994/01 to 2000/12 is presented in table 16 to table 17. In table 16, we can see that for 6-month ahead forecasting, LSTM model performs slightly better than both random walk and the DNS model except for the 3-month maturity one. To check if the forecasts of LSTM are statistically significantly better, I conduct the Diebold-Mariano test (26) in table 18. The Diebold-Mariano test shows that in some maturities, LSTM indeed performs better than the DNS model and random walk. For 12-month ahead forecasting, LSTM performs better than random walk and the DNS model in all maturities, and the test in table 18 also confirms the good performances of LSTM are statistically significant. These results are consistent with what we see in [25] for the DNS model. Moreover, the results show that LSTM performs better than the DNS model in 1994-2000.

In table 19, we see the results for 6-month ahead forecasting from 2015/01 to 2019/12. For most maturities, out-of-sample RMSEs for LSTM are worse than random walk. On the other hand, the DNS model is obviously worse than the other two models for most maturities. This result is very different from previous findings in [25]. In their paper, they showed that the DNS model has significantly better out-of-sample forecasts than random walk in 1994/01

Table 16: Out-of-sample 6-month-ahead Forecasting Errors

Maturity(τ)	Mean	Std. Dev.	RMSE
Random Walk			
3 months	-0.2060	0.5507	0.5880
12 months	-0.1980	0.7241	0.7506
36 months	-0.1283	0.8456	0.8553
60 months	-0.0790	0.8237	0.8274
120 months	-0.0161	0.7166	0.7168
DNS			
3 months	-0.0224	0.7973	0.7976
12 months	-0.3252	0.8195	0.8816
36 months	-0.2294	0.8180	0.8496
60 months	0.0178	0.7823	0.7825
120 months	0.1905	0.6860	0.7120
LSTM with attention encoder and decoder			
3 months	-0.2259	0.6148	0.6550
12 months	-0.3097	0.6707	0.7387
36 months	-0.1117	0.7263	0.7348
60 months	0.0256	0.7278	0.7282
120 months	0.0838	0.6877	0.6928

Table 17: Out-of-sample 12-month-ahead Forecasting Results

Maturity(τ)	Mean	Std. Dev.	RMSE
Random Walk			
3 months	-0.3880	0.9005	0.9804
12 months	-0.3872	1.0876	1.1544
36 months	-0.2550	1.1940	1.2209
60 months	-0.1533	1.1483	1.1585
120 months	0.0128	1.0070	1.0071
DNS			
3 months	-0.2060	0.9982	1.0192
12 months	-0.4815	1.0032	1.1128
36 months	-0.2260	0.9772	1.0030
60 months	0.1229	0.8972	0.9055
120 months	0.9901	0.7557	0.9136
LSTM with encoder and decoder			
3 months	-0.3191	0.8243	0.8840
12 months	-0.4042	0.8495	0.9407
36 months	-0.2108	0.8734	0.8984
60 months	-0.0770	0.8707	0.8741
120 months	-0.0232	0.8455	0.8458

to 2000/12. However, it is not the case here in 2015/01 to 2019/12. On the other hand, although LSTM does not forecast better than random walk in most maturities, it performs much better than DNS. In table 21, the Diebold and Mariano test also confirms that random walk performs the best and the DNS model forecast the worst for 6-month ahead forecasting in 2015 to 2019.

Table 18: Out-of-sample Forecast Accuracy Comparisons

Maturity(τ)	Against RW (6-month horizon)	Against DNS (6-month horizon)	Against RW (12-month horizon)	Against DNS (12-month horizon)
3 months	1.081	-2.603*	-1.423	-1.906*
12 months	-0.171	-2.372*	-2.796*	-2.775*
36 months	-1.796*	-2.280*	-4.485*	-1.568
60 months	-1.747*	-1.410	-5.034*	-0.419
120 months	-0.519	-0.492	-3.499*	-0.721

I present Diebold–Mariano forecast accuracy comparison tests of the LSTM model forecasts against those of the random walk model (RW) and the dynamic Nelson Siegel model (DNS). The null hypothesis is that the two forecasts have the same mean squared error. Negative values indicate superiority of the LSTM model forecasts, and asterisks denote significance relative to the asymptotic null distribution at the 10 percent level.

For 12-month-ahead forecasting results in table 20, LSTM preforms better than both random walk and DNS model in all maturities in terms of RMSEs. When checking the Diebold-Mariano test results in table 21, LSTM is statistically significantly better than random walk in 3-month, 36-month and 120-month forecasts. For the test against DNS, LSTM is also significantly better than DNS in all maturities.

Table 19: Out-of-sample 6-month-ahead Forecasting Errors

Maturity(τ)	Mean	Std. Dev.	RMSE
Random Walk			
3 months	-0.1793	0.3142	0.3617
12 months	-0.1588	0.3599	0.3934
36 months	-0.0607	0.4484	0.4525
60 months	0.0017	0.4794	0.4794
120 months	0.0588	0.5005	0.5039
DNS			
3 months	-0.0752	0.8415	0.8449
12 months	-0.2194	0.8814	0.9083
36 months	-0.0024	0.8547	0.8547
60 months	0.1426	0.7713	0.7843
120 months	0.2800	0.6577	0.7149
LSTM with attention encoder and decoder			
3 months	0.0312	0.3479	0.3492
12 months	0.0096	0.4508	0.4509
36 months	0.1587	0.5431	0.5658
60 months	0.1375	0.5617	0.5783
120 months	0.0289	0.5637	0.5644

Table 20: Out-of-sample 12-month-ahead Forecasting Errors

Maturity(τ)	Mean	Std. Dev.	RMSE
Random Walk			
3 months	-0.4101	0.4301	0.5943
12 months	-0.3860	0.5068	0.6371
36 months	-0.2057	0.6248	0.6578
60 months	-0.0682	0.6264	0.6301
120 months	0.0774	0.6211	0.6259
DNS			
3 months	-0.1816	0.6922	0.7157
12 months	-0.4377	0.7422	0.8617
36 months	-0.1871	0.8905	0.9100
60 months	0.0814	0.8494	0.8533
120 months	0.9709	0.7679	0.8432
LSTM with encoder and decoder			
3 months	-0.3026	0.4725	0.5611
12 months	-0.2864	0.5543	0.6239
36 months	-0.0464	0.5907	0.5925
60 months	-0.0074	0.5689	0.5689
120 months	-0.0471	0.5462	0.5483

Table 21: Out-of-sample Forecast Accuracy Comparisons

Maturity(τ)	Against RW (6-month horizon)	Against DNS (6-month horizon)	Against RW (12-month horizon)	Against DNS (12-month horizon)
3 months	-0.697	-4.294*	-1.962*	-4.993*
12 months	2.816*	-5.239*	-0.516	-7.177*
36 months	3.098*	-4.284*	-1.672*	-6.678*
60 months	2.556*	-3.626*	-1.424	-4.035*
120 months	2.054*	-2.601*	-1.940*	-3.041*

I present Diebold–Mariano forecast accuracy comparison tests of the LSTM model forecasts against those of the random walk model (RW) and the dynamic Nelson Siegel model (DNS). The null hypothesis is that the two forecasts have the same mean squared error. Negative values indicate superiority of the LSTM model forecasts, and asterisks denote significance relative to the asymptotic null distribution at the 10 percent level.

3.5 Discussion and Conclusion

From the out-of-sample forecasting, we can see the possibility of having a better forecasting accuracy in terms of RMSE by using recurrent neural networks, especially for the forecasts of longer horizons. Although the forecast results show that LSTM may not always performs better than random walk, it usually gives more precise forecasts comparing to the DNS model. In addition, the forecast from LSTM still inherits most of the theoretical properties from the DNS model. Moreover, since the model I propose in this paper is a generalization of the original DNS model, imposing the no-arbitrage conditions in the previous literature (19) on my model is also feasible. The reasons above make LSTM a preferred option when doing DNS-type of yield curve forecasting.

The proposed forecasting model in this paper can be seen as a "hybrid" model, which incorporates the existing theoretical frameworks in financial forecasting with models in machine learning, rather than just using pure machine learning methods. The methodology of using a mixture of traditional models and machine learning is also prospering in many other fields and is proven to give better performance in out-of-sample forecasting (49).

To sum up, this paper shows how recurrent neural network can carry similar theoretical properties from previous financial econometrics literature while improving the forecasting accuracy. Further understanding about how we can incorporate other desirable properties from financial theories with this model can help us link it closer to traditional financial theories.

4.0 Appendix

4.1 Attribution Bias on Online Reputation Systems

4.1.1 Graphs for Regression Discontinuity

Table 22: Summary Statistics for All Users

	Obs	Mean	Std Dev	Min	Max
Reviewer's Rating	20897	3.839690	1.367365	1	5
Number of Fans	20897	10.964588	55.844576	0	2383
Written By an Elite or Not	20897	0.18778	0.390544	0	1
Good feedback for a specific review	20897	2.144040	8.234336	0	565
Total Reviews for a Restaurant	20897	369.869694	377.851384	101	3415

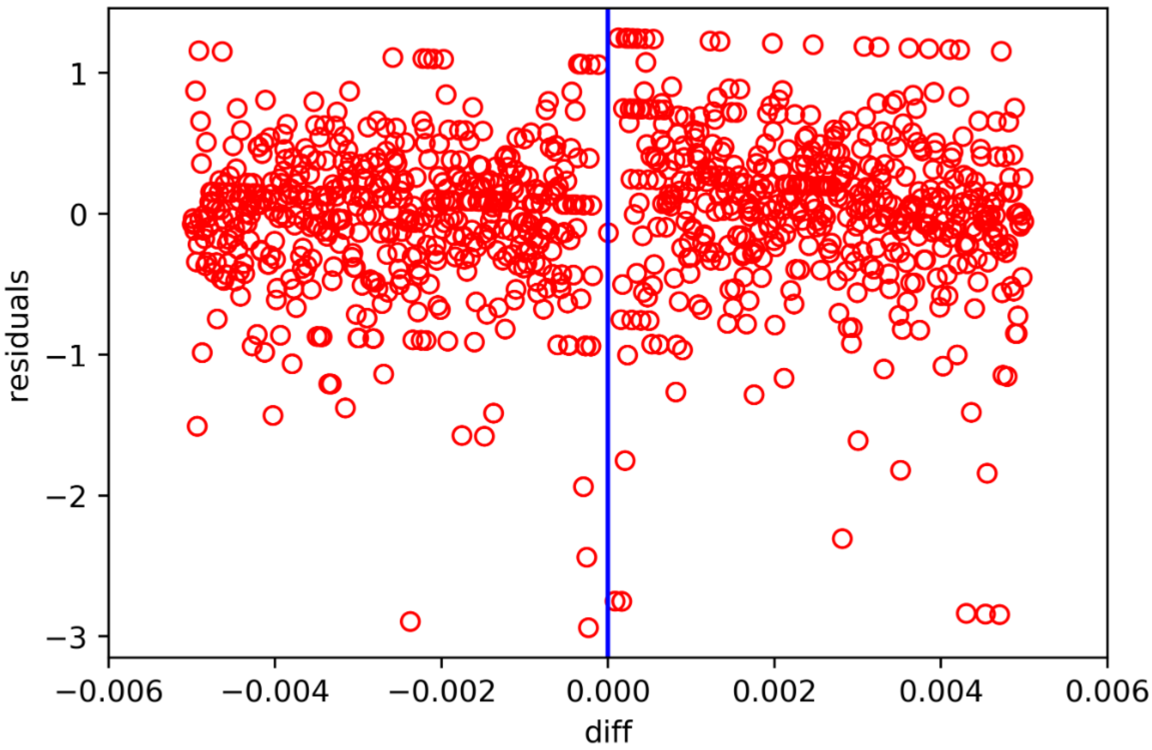


Figure 10: Residual Plot of Equation (5)

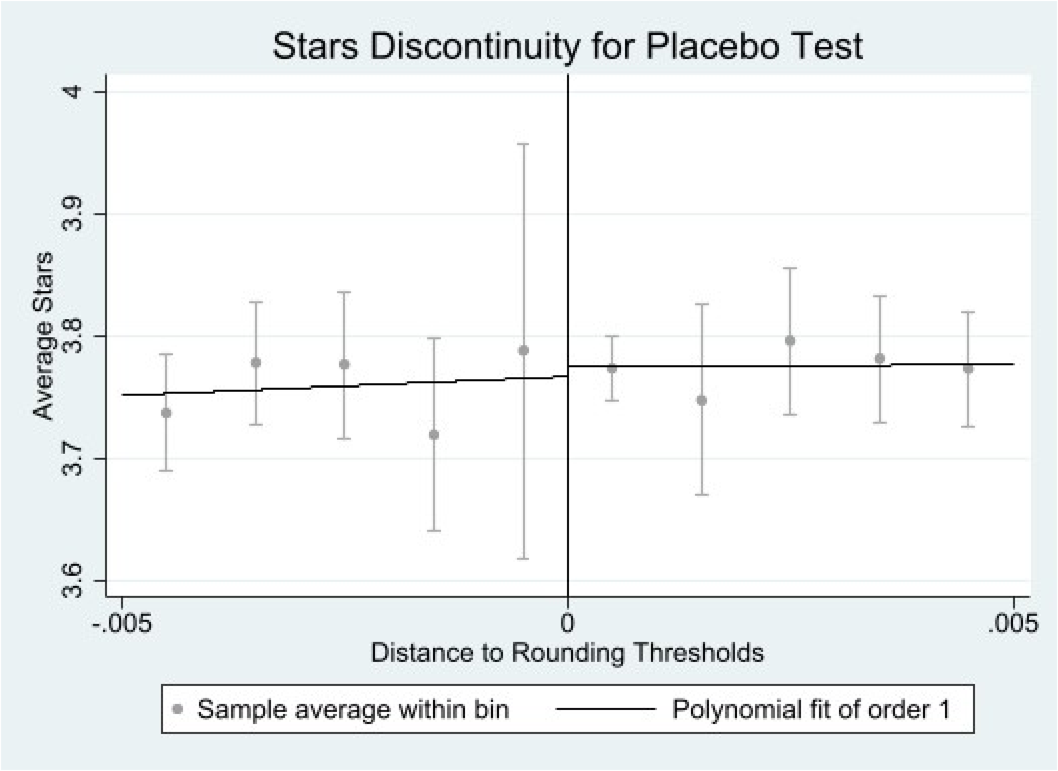


Figure 11: Discontinuity Plot for the Placebo Test

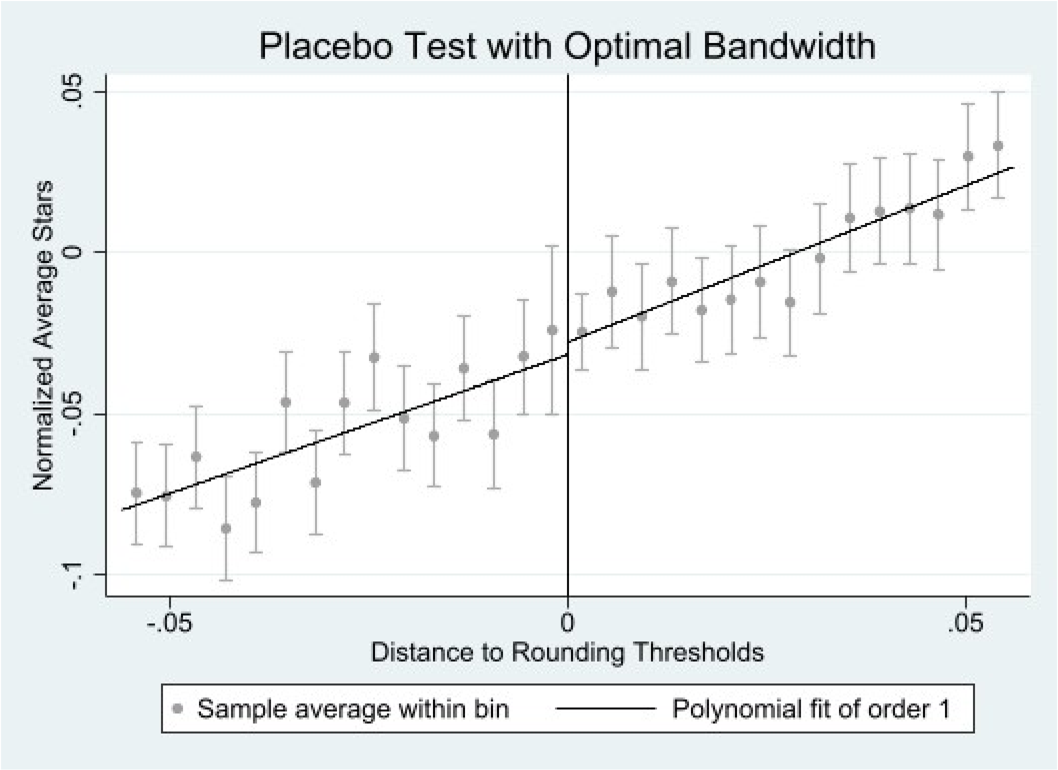


Figure 12: Discontinuity Plot of Normalized Ratings with the Optimal Bandwidth

4.1.2 More Robustness Checks

Table 23: Ordered Probit Regression Results of Reference-Dependent Behaviors

	(1) Reviewer's Rating	(2) Reviewer's Rating	(3) Reviewer's Rating	(4) Reviewer's Rating
RD	-0.5399***	-0.0162	-0.4858***	-0.046
	(0.0342)	(0.024)	(0.076)	(0.0533)
diff	7.023302	4.2489	-1.57811	11.0434
	(5.3853)	(3.934)	(11.9684)	(8.7209)
Controls	Yes	Yes	Yes	Yes
Observations	20897	30955	3924	5888

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the estimation with all users

(2) is the placebo test for (1) with imaginary rounding thresholds

(3) includes only elite users.

(4) is the placebo test for (3) with imaginary rounding thresholds

Table 24: Ordered Probit Model with the Optimal Bandwidth (0.056)

	(1)	(2)	(3)	(4)
	Reviewer's Rating	Reviewer's Rating	Reviewer's Rating	Reviewer's Rating
RD	-0.4748***	0.0001	-0.4451***	0.0059
	(0.0061)	(0.0052)	(0.0137)	(0.012)
diff	1.3834***	0.8421***	1.2576***	0.7636***
	(0.0894)	(0.0814)	(0.1992)	(0.1806)
Controls	Yes	Yes	Yes	Yes
Observations	651775	714161	123679	135724

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the estimation with all users

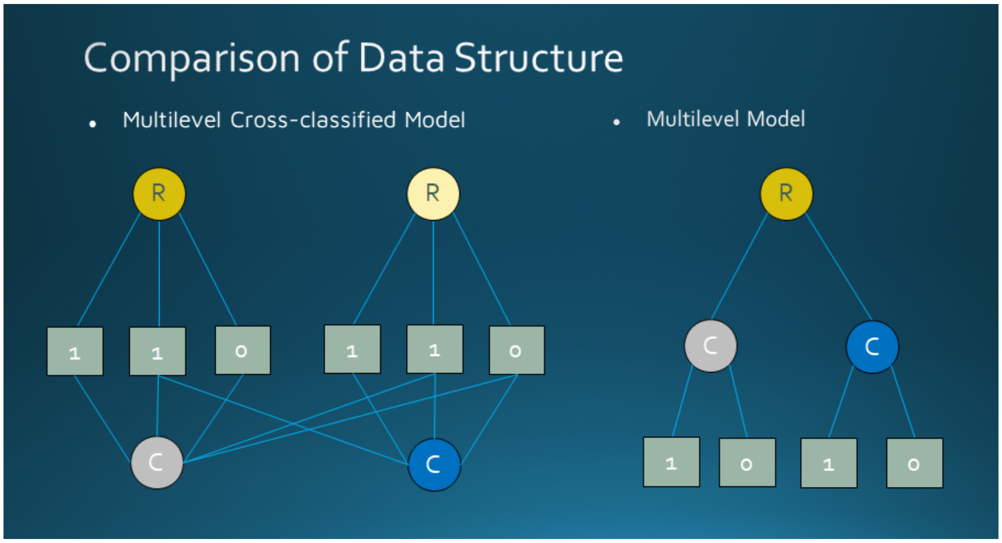
(2) is the placebo test for (1) with imaginary rounding thresholds

(3) includes only elite users.

(4) is the placebo test for (3) with imaginary rounding thresholds

4.2 Hope Hurts: Attribution Bias in Yelp Reviews

4.2.1 Cross-Classified Multilevel Models



Comparison of Data Structure

- Multilevel Cross-classified Model
- Multilevel Model

Figure 13: Comparison between Cross-Classified Multilevel Model and Multilevel Model

In this figure, R represents restaurants, and C stands for consumers. A square with number 1 shows that a consumer went to the restaurant on a special occasion. A 0 means that a consumer went there on a non-special occasion. In cross-classified multilevel models, consumers may write reviews for several restaurants on different occasions, and there is no clear hierachy between restaurants and consumers. In normal multilevel models, customers only write one review for one restaurant and the restaurant includes all the reviews of those customers.

4.2.2 Distribution of Elites/Non-Elites

Table 25: The Distributions of Ratings of Elites/Non Elites

	(1)	(2)	(3)	(4)	(5)
	1 star	2 star	3 star	4 star	5 star
Non-elite in the sample	13.93%	8.64%	8.92%	17.77%	50.75%
Yelp elite in the sample	4.73%	6.48%	14.12%	32.00%	42.67%
Population Distribution	17%	7%	8%	17%	51%

This table shows the elite/non elite rating distributions of the sample of our analyses and the population distribution on Yelp

4.2.3 Honest Causal Tree

This subsection shows the estimation results of the heterogeneous treatment effects for each subpopulation by causal tree algorithm (3). In order to avoid overfitting, which means the calculated estimation may not be generalized to the population, we perform the honest causal tree algorithm (3). The honest approach divides the data into two parts, the splitting subsample and the estimating subsample. The splitting subsample is used to construct the partition and build a causal tree and the estimating subsample to estimate unbiased treatment effects for each subpopulation. In this algorithm, we build a causal tree to minimize $-\tau(x;\Pi)^2$, where $\tau(x;\Pi) = E(Y(1)-Y(0)|x) \in l(x;\Pi))$. Here $Y(1)-Y(0)$ is the treatment effect, Π is a tree partition and $l(x;\Pi)$ denotes the leaf $l \in \Pi$ such that $x \in l$.

The pruned honest casual tree is displayed below. Obviously, average stars, reference point, and whether the user belongs to the Yelp Elite all influence the treatment heterogeneity. This output resembles our previous results generated by econometric models. We also take a closer look at the heterogeneous treatment effects for Yelp elites/non-elites. When the user is not a Yelp elite and the average star of the restaurant is below 4.4, the effect of negative disconfirmation (of expectation) is relatively strong and statically significant. However, if a user belongs to the Yelp elite squad, the effect of negative disconfirmation becomes much smaller, and is non-significant. The treatment heterogeneity of Yelp elites is detected again by this method: compared with non-elites, these Yelp elites suffer less from this cognitive bias.

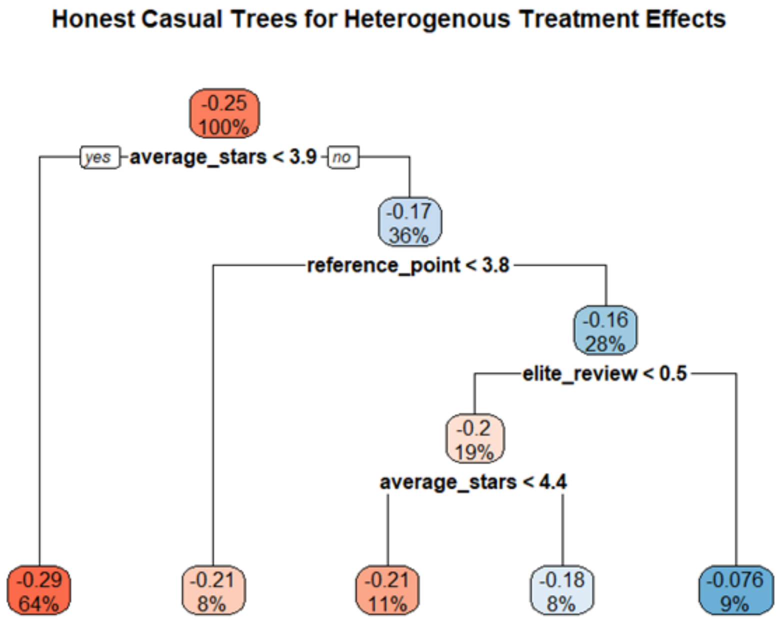


Figure 14: Results of Honest Causal Tree

Table 26: Casual Tree Estimation Results

	(1) leaf 1	(2) leaf 2	(3) leaf 3	(4) leaf 4	(5) leaf 5
Treatment effect of Special Occasion	-0.2936	-0.2145	-0.2137	-0.1829	-0.0761
	(0.027)	(0.066)	(0.078)	(0.076)	(0.069)

This table shows the treatment effects of 5 terminal leaves from left to right and standard errors are in the parenthesis

4.2.4 Censored Least Absolute Deviations Estimators

Table 27: Results of Special Occasion Effect on *Repeated Reviews* for Tobit Models

	(1) Reviewer's rating
SpecialOccasion	-0.2489***
	(0.011)
Controls	Yes
Observations	12632

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the CLAD result with controls for business average ratings

, reference points and visit times

4.2.5 Ordered Logit Estimation Results

Table 28: Ordered Logit Results of Special Occasion Effect on *Repeated Reviews*

	(1) Reviewer's rating	(2) Reviewer's rating	(3) Reviewer's rating	(4) Reviewer's rating
$SpecialOccasion$	-0.2354***	-0.2672***	-0.1943***	-0.1782***
	(0.033)	(0.033)	(0.014)	(0.045)
Observations	12632	12632	12632	12632

standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(1) presents the OLS result without any controls

(2) shows the OLS result with fixed effects for different rating ranges

(3) shows the OLS result with column 2's fixed effect and controls for repeated visits

(4) adds potential reference points and more control variables

4.2.6 Balance Check for Propensity Score Matching

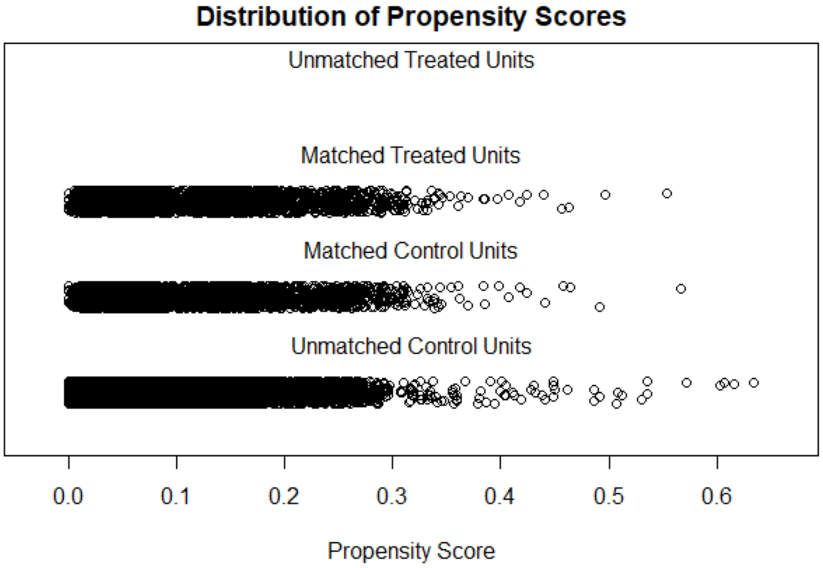


Figure 15: Distribution of Propensity Scores

4.3 From Econometrics to Machine Learning: Application of Recurrent Neural Networks on Yield Curve Forecasting

4.3.1 Dual-Stage Attention-Based Encoder and Decoder

Attention mechanism is a special structure which puts extra emphases on certain inputs or particular parts of a neural network by reweighting the coefficient of those elements. It is used more and more often in machine learning, especially in fields like image captioning (86) and document classification, to improve the performance of models.

In [76], the authors proposed a dual-stage encoder-decoder model to improve the time-series forecasting performance of nonlinear autoregressive exogenous models. Inspired by theories of human attention in psychology, they added two attention mechanisms, input attention and temporal attention, in a encoder-decoder structure and showed that their dual-

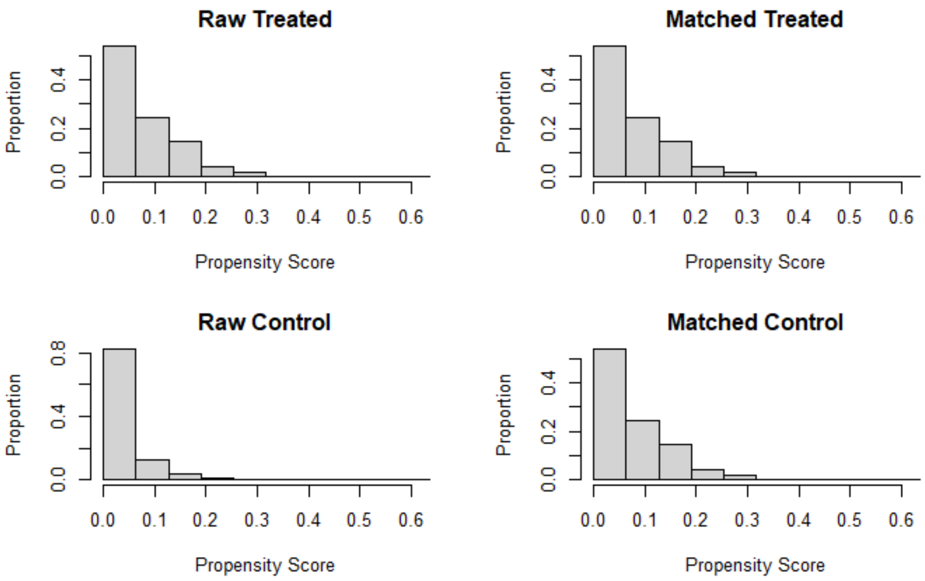


Figure 16: Distributions Before and After Matching

stage attention based encoder-decoder network peformed better than traditional nonlinear autoregressive exogenous models.

Following their work, I extend the dual-stage attention based encoder and decoder to incorporate the case with no exogenous varialbe in my forecasting model. Moreover, the extension allows me to forecast multidimensional variables rather than just univaraite cases in the original paper.

The dual-stage attention based encoder-decoder network is based on the LSTM model I show in section 3.3.2.2: Considering two LSTMs and let one be an encoder and the other be an decoder.

In the encoder, the hidden state, h_t, and output vector, O_t, evolve in the following way:

$$h_t = f_1(h_{t-1}, x_{\tilde{t-1}})$$

$$O_t = g_1(h_t)$$

g_1 is a differentiable activation function and the structure of the function f_1 is:

$$\mathfrak{f}_t = \sigma(W_{\mathfrak{f}} x_{\tilde{t-1}} + U_f h_{t-1} + b_f)$$

$$
\begin{aligned}
i_t &= \sigma(W_i \tilde{x_{t-1}} + U_i h_{t-1} + b_i) \\
\mathfrak{o}_t &= \sigma(W_\mathfrak{o} \tilde{x_{t-1}} + U_\mathfrak{o} h_{t-1} + b_\mathfrak{o}) \\
\tilde{c_t} &= tanh(W_c \tilde{x_{t-1}} + U_c h_{t-1} + b_c) \\
c_t &= f_t \otimes c_{t-1} + i_t \otimes \tilde{c_t} \\
h_t &= \mathfrak{o}_t \otimes tanh(c_t)
\end{aligned}
$$

Here, $W_\mathfrak{f}$, W_i, $W_\mathfrak{o}$, $W_c \in \mathbb{R}^{h \times d}$ and U_f, U_i, $U_\mathfrak{o}$, $U_c \in \mathbb{R}^{h \times h}$ are the coefficient matrices. b_f, b_i, $b_\mathfrak{o}$, $b_c \in \mathbb{R}^h$ are constant vectors. Moreover,

$\tilde{x_t} \in R^d$: reweighted input by the input attention mechanism

$\mathfrak{f}_t \in R^h$: forget gate's activation vector

$i_t \in R^h$: input/update gate's activation vector

$\mathfrak{o}_t \in R^h$: output gate's activation vector

$h_t \in R^h$: hidden state vector

$\tilde{c_t} \in R^h$: cell input activation vector

$c_t \in R^h$: cell state vector

The input attention mechanism works on the original input, x_t, and transforms x_t to $\tilde{x_t}$ in the following way: Let the k-th component of the input series $x^k = (x_1^k, x_2^k, ..., x_T^k)^\intercal \in \mathbb{R}^T$ and let

$$
\begin{aligned}
e_t^k &= v_e^\intercal tanh(W_e[h_{t-1}; c_{t-1} + U_e x^k] + b_e) \\
\alpha_t^k &= \frac{exp(e_t^k)}{\sum_{i=1}^{d} exp(e_t^i)}
\end{aligned}
$$

where $v_e \in \mathbb{R}^T$, $W_e \in R^{T \times 2h}$, and $U_e \in \mathbb{R}^{T \times T}$. Here α_t^k works as a new weight for x_t^k and the transformed $\tilde{x_t} = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, ..., \alpha_t^n x_t^n)^\intercal$. After the input attention mechanism, we put $\tilde{x_t}$ back to the LSTM and get the hidden-state vector $H = \{h_1, h_2, ..., h_{T+1}\}$ from the encoder.

After the encoder, the hidden-state vector H is fed into the decoder. The decoder is another LSTM network with a temporal attention mechanism. In the decoder, the hidden state, h'_t, and output vector, O'_t, evolve similarly as the encoder:

$$\begin{aligned} h'_t &= f_2(h'_{t-1}, \tilde{y_{t-1}}) \\ O'_t &= g_2(h'_t) \end{aligned}$$

g_2 is a differentiable activation function and the structure of the function f_2 is:

$$\begin{aligned} \mathfrak{f}'_t &= \sigma(W_{\mathfrak{f}'} \tilde{y_{t-1}} + U'_f h'_{t-1} + b'_f) \\ i'_t &= \sigma(W'_i \tilde{y_{t-1}} + U'_i h'_{t-1} + b'_i) \\ \mathfrak{o}'_t &= \sigma(W'_{\mathfrak{o}} \tilde{y_{t-1}} + U'_{\mathfrak{o}} h'_{t-1} + b'_{\mathfrak{o}}) \\ \tilde{c_t}' &= tanh(W'_c \tilde{y_{t-1}} + U'_c h'_{t-1} + b'_c) \\ c'_t &= f'_t \otimes c_{t-1} + i'_t \otimes \tilde{c_t}' \\ h'_t &= \mathfrak{o}'_t \otimes tanh(c'_t) \end{aligned}$$

Here, $W'_{\mathfrak{f}}$, W'_i, $W'_{\mathfrak{o}}$, $W'_c \in \mathbb{R}^{p\times 3}$ and U'_f, U'_i, $U'_{\mathfrak{o}}$, $U'_c \in \mathbb{R}^{p\times p}$ are the coefficient matrices. b'_f, b'_i, $b'_{\mathfrak{o}}$, $b'_c \in \mathbb{R}^p$ are constant vectors. Moreover,

$y_t \in R^3$: $\hat{\beta}$s from the estimation of equation (1)

$\tilde{y}_t \in R^3$: reweighted $\hat{\beta}$s by the temporal attention mechanism

$\mathfrak{f}'_t \in R^p$: forget gate's activation vector

$i'_t \in R^p$: input/update gate's activation vector

$\mathfrak{o}'_t \in R^p$: output gate's activation vector

$h'_t \in R^p$: hidden state vector

$\tilde{c_t}' \in R^p$: cell input activation vector

$c'_t \in R^p$: cell state vector

By taking the h_ts from the encoder with hidden states and cell states from the decoder, the decoder uses a temporal attention mechanism to create coefficients for reweighting h_ts. Let

$$
\begin{aligned}
&l_t^i = v_{h'}^{\intercal} tanh(W_{h'}[h'_{t-1}; c'_{t-1} + U_{h'}h_i] + b_{h'}),\ 1 \leq i \leq T+1 \\
&\gamma_t^i = \frac{exp(l_t^i)}{\sum_{j=1}^{T+1} exp(l_t^j)} \\
&W_{h'} \in \mathbb{R}^{h \times (2p)},\ v_{h'} \in \mathbb{R}^{h},\ U_{h'} \in \mathbb{R}^{h \times h} \text{ and } b_{h'} \in \mathbb{R}^{h}
\end{aligned}
$$

Using the γ_t^i to reweight h_i, we can get a context vector s_t

$$
s_t = \sum_{i=1}^{T+1} \gamma_t^i h_i
$$

Continuing with the context vector, we can get a reweighted y_{t-1} as

$$
\begin{aligned}
&\tilde{y_{t-1}} = \tilde{w}[y_{t-1}; s_{t-1}] + \tilde{b} \\
&\tilde{w} \in \mathbb{R}^{3 \times (h+3)} \text{ and } \tilde{b} \in \mathbb{R}^{3}
\end{aligned}
$$

Finally, we use the hidden state of the decoder, $\tilde{h_{T+1}}$ and s_{T+1} and get the forecast $\hat{y_{T+1}}$ as

$$
\begin{aligned}
&\hat{y_{T+1}} = v_y(W_y[h'_{T+1}; s_{T+1}] + b_w) + b_v \\
&W_y \in \mathbb{R}^{p \times (p+h)},\ v_y \in \mathbb{R}^{3 \times p},\ b_w \in \mathbb{R}^{p},\ b_v \in \mathbb{R}^{3}
\end{aligned}
$$

4.3.2 Graph for Forecasting Errors

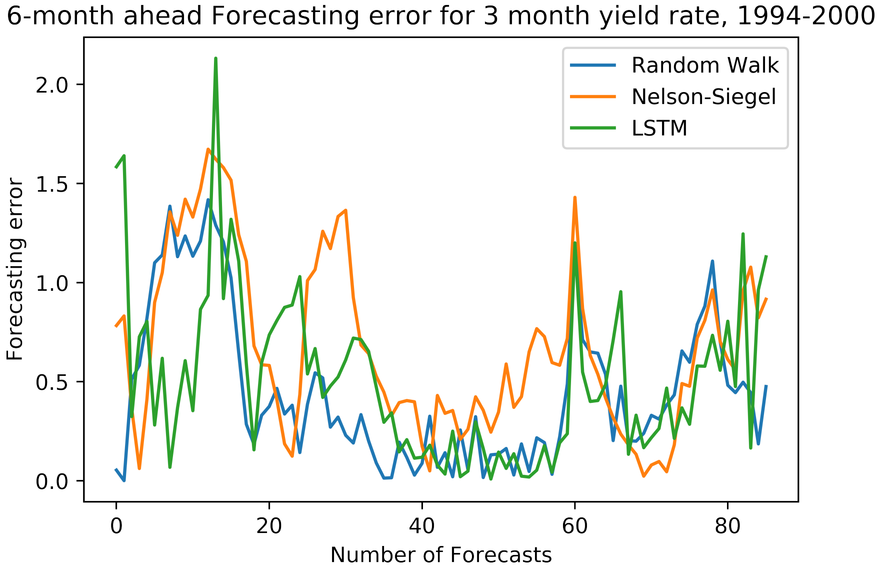


Figure 17: 6-month Ahead Forecast Errors of Each Period for 3-month Maturity from 1994 to 2000

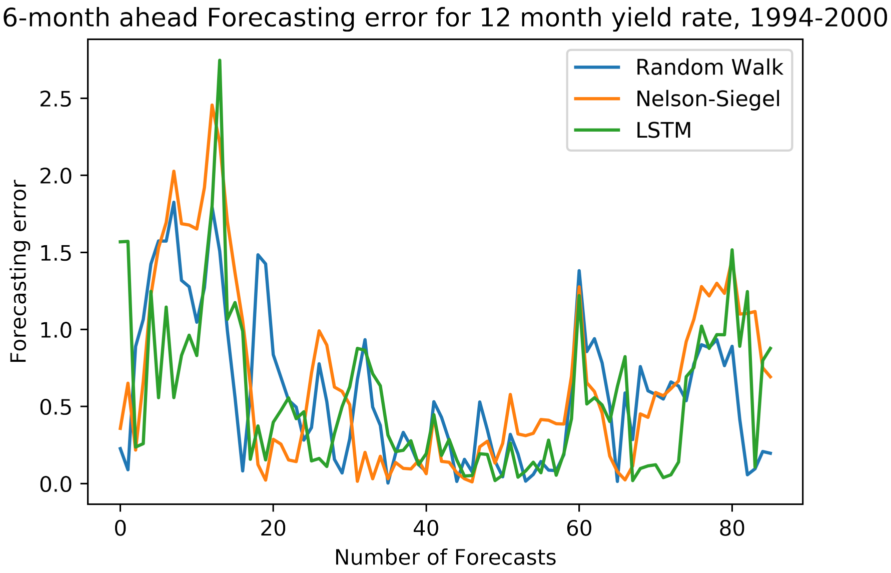


Figure 18: 6-month Ahead Forecast Errors of Each Period for 12-month Maturity from 1994 to 2000

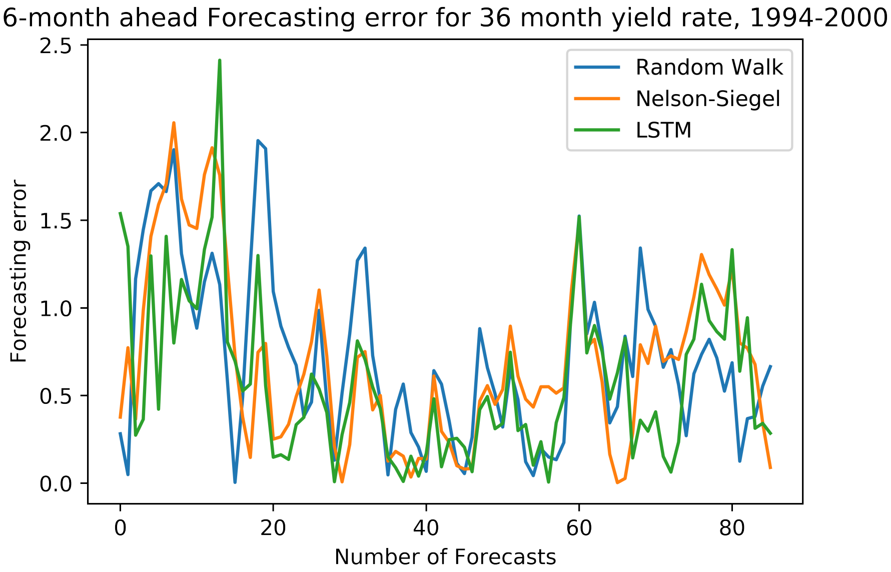


Figure 19: 6-month Ahead Forecast Errors of Each Period for 36-month Maturity from 1994 to 2000

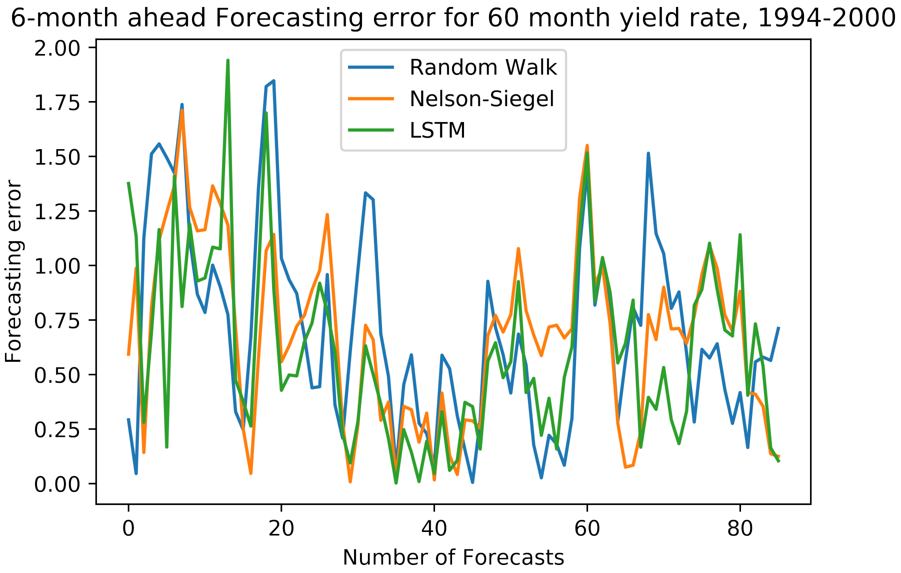


Figure 20: 6-month Ahead Forecast Errors of Each Period for 60-month Maturity from 1994 to 2000

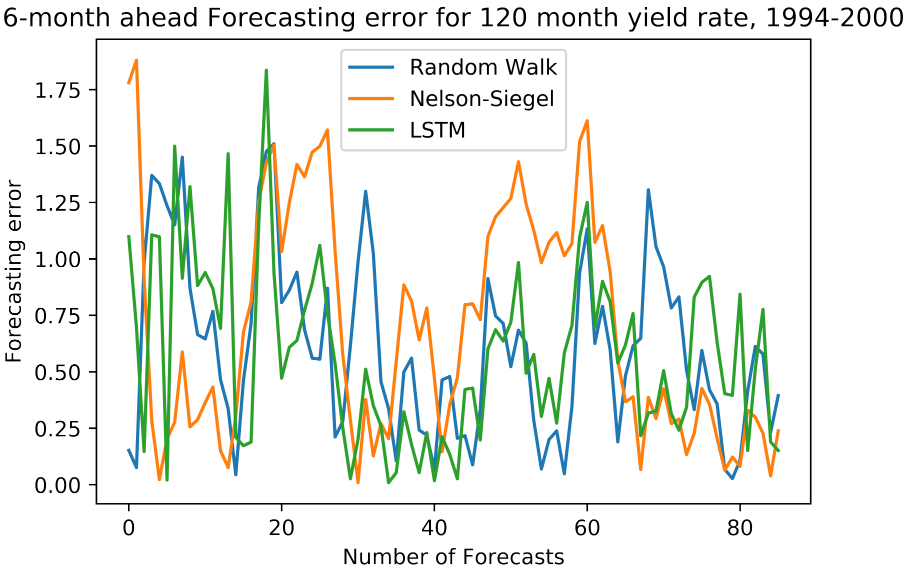


Figure 21: 6-month Ahead Forecast Errors of Each Period for 120-month Maturity from 1994 to 2000

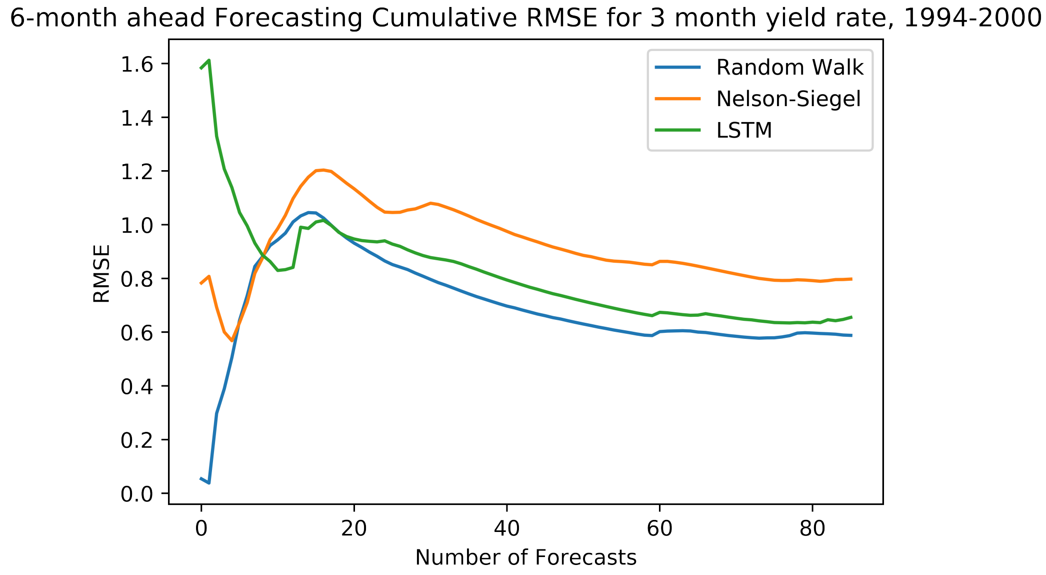


Figure 22: 6-month Ahead Forecast RMSEs for 3-month Maturity from 1994 to 2000

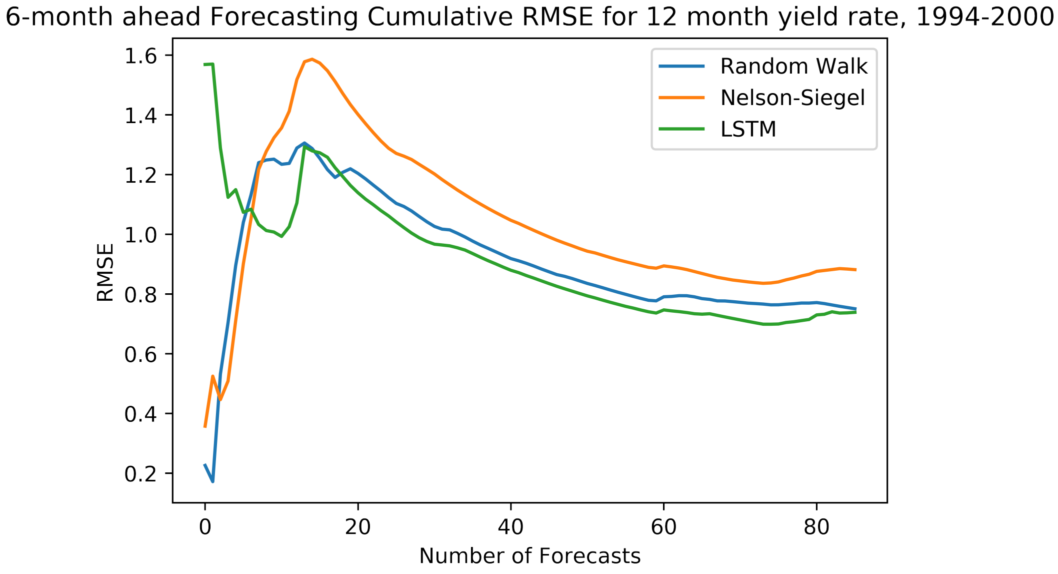


Figure 23: 6-month Ahead Forecast RMSEs for 12-month Maturity from 1994 to 2000

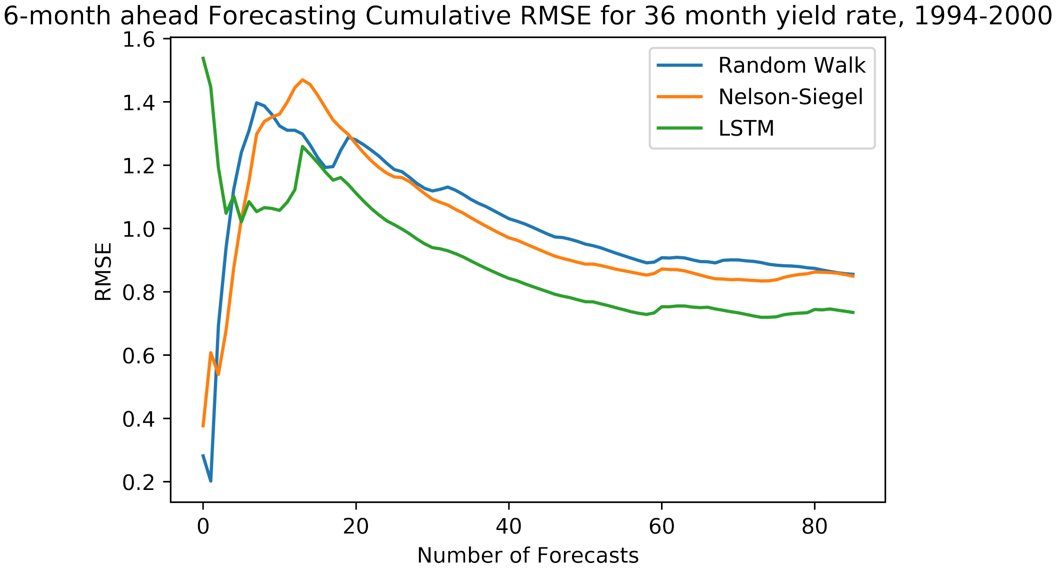


Figure 24: 6-month Ahead Forecast RMSEs for 36-month Maturity from 1994 to 2000

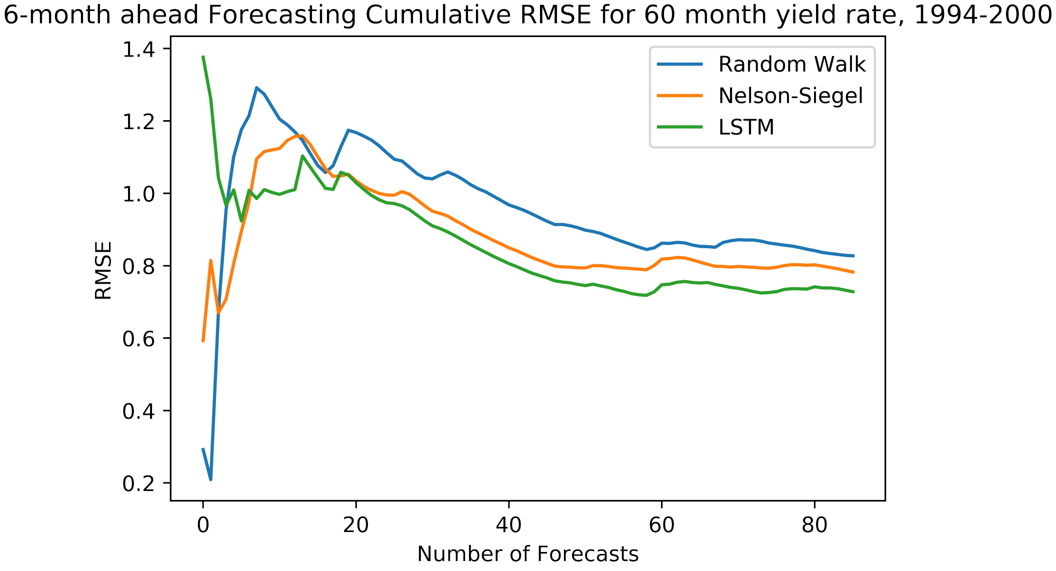


Figure 25: 6-month Ahead Forecast RMSEs for 60-month Maturity from 1994 to 2000

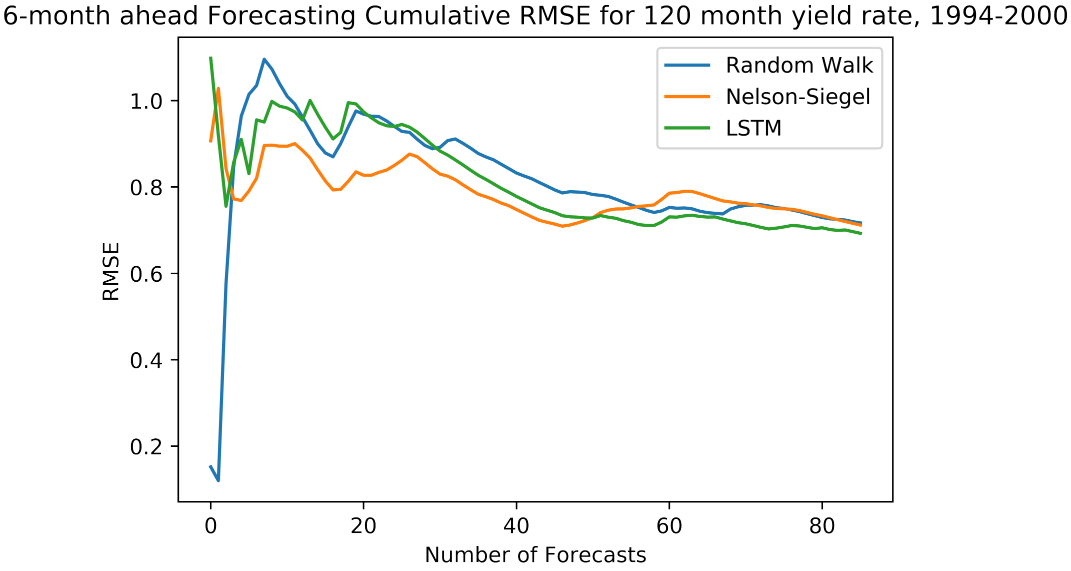


Figure 26: 6-month Ahead Forecast RMSEs for 120-month Maturity from 1994 to 2000

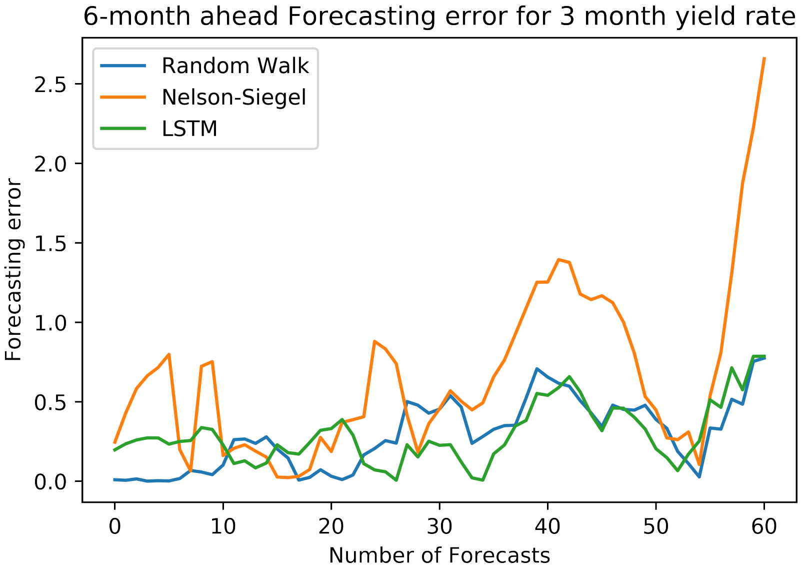


Figure 27: 6-month Ahead Forecast Errors of Each Period for 3-month Maturity from 2015 to 2019

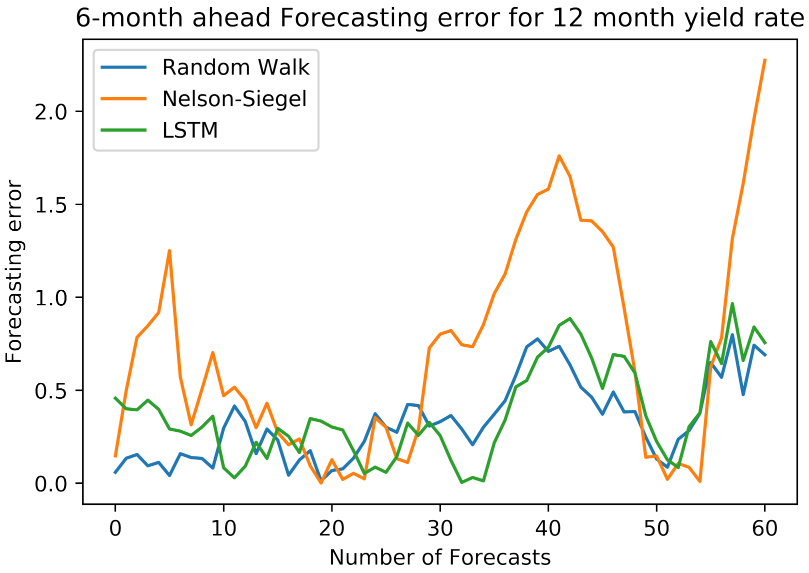


Figure 28: 6-month Ahead Forecast Errors of Each Period for 12-month Maturity from 2015 to 2019

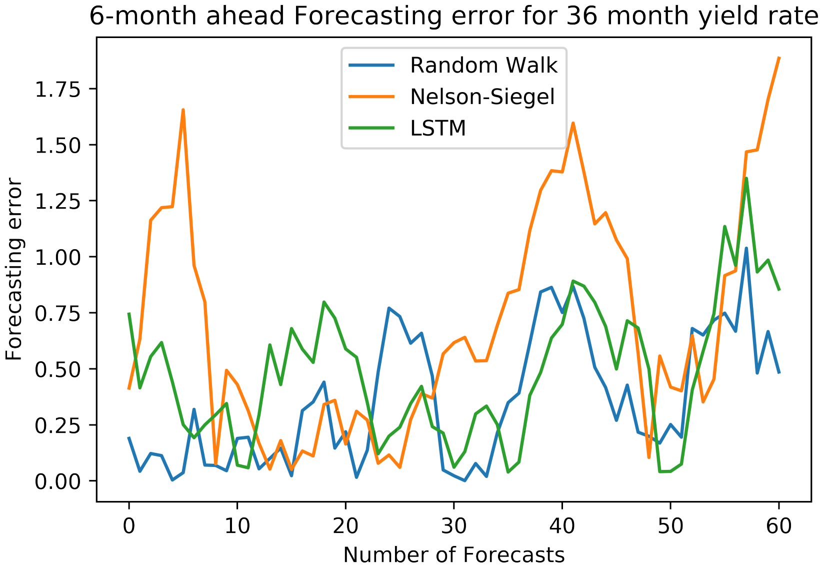


Figure 29: 6-month Ahead Forecast Errors of Each Period for 36-month Maturity from 2015 to 2019

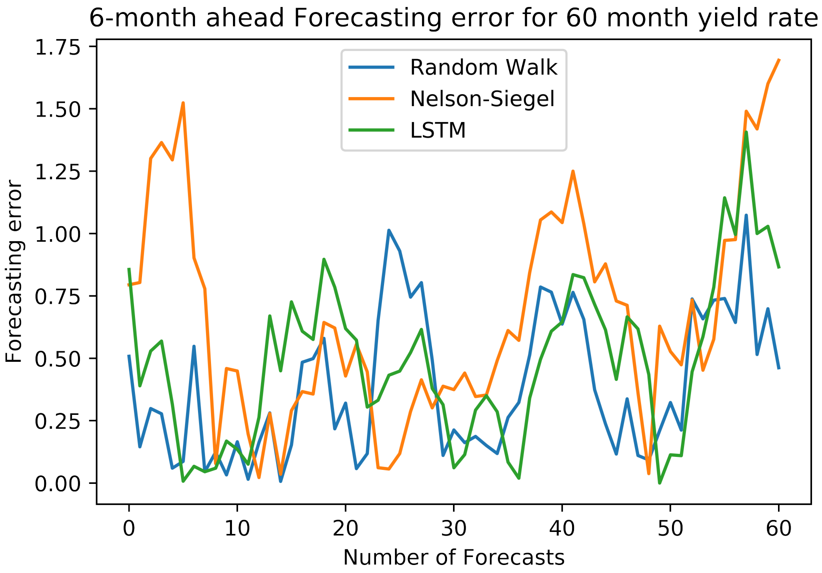


Figure 30: 6-month Ahead Forecast Errors of Each Period for 60-month Maturity from 2015 to 2019

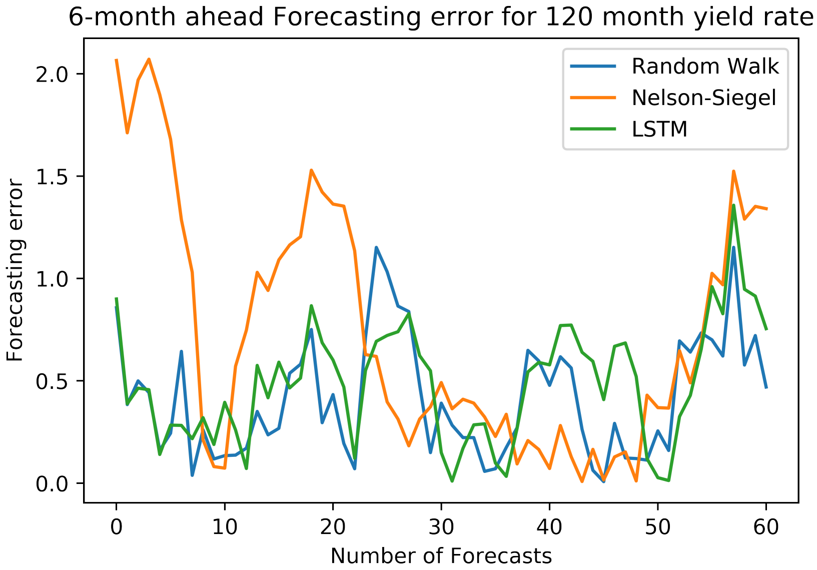


Figure 31: 6-month Ahead Forecast Errors of Each Period for 120-month Maturity from 2015 to 2019

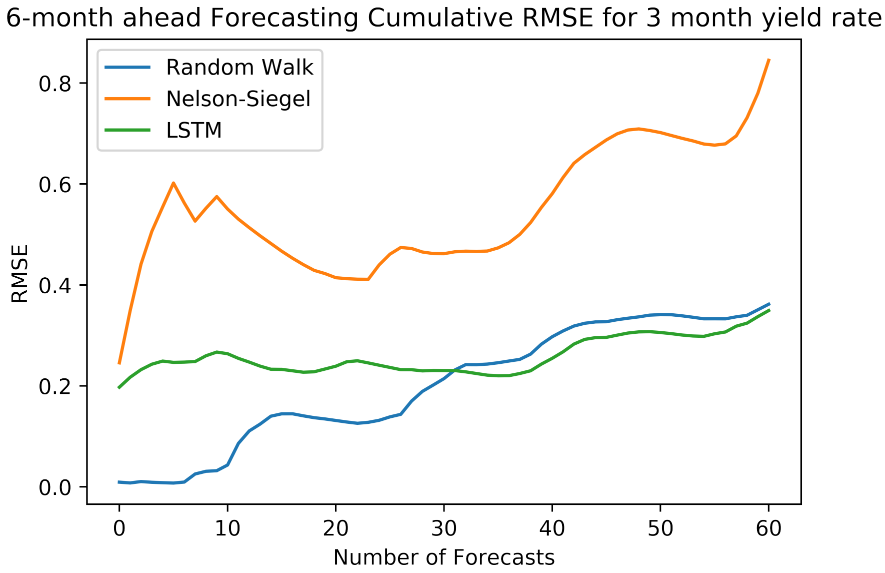


Figure 32: 6-month Ahead Forecast RMSEs for 3-month Maturity from 2015 to 2019

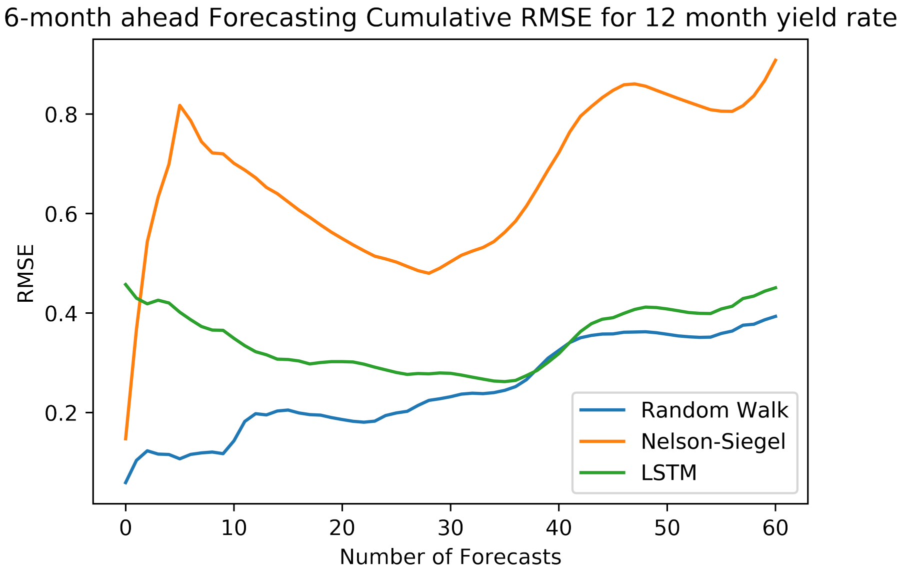


Figure 33: 6-month Ahead Forecast RMSEs for 12-month Maturity from 2015 to 2019

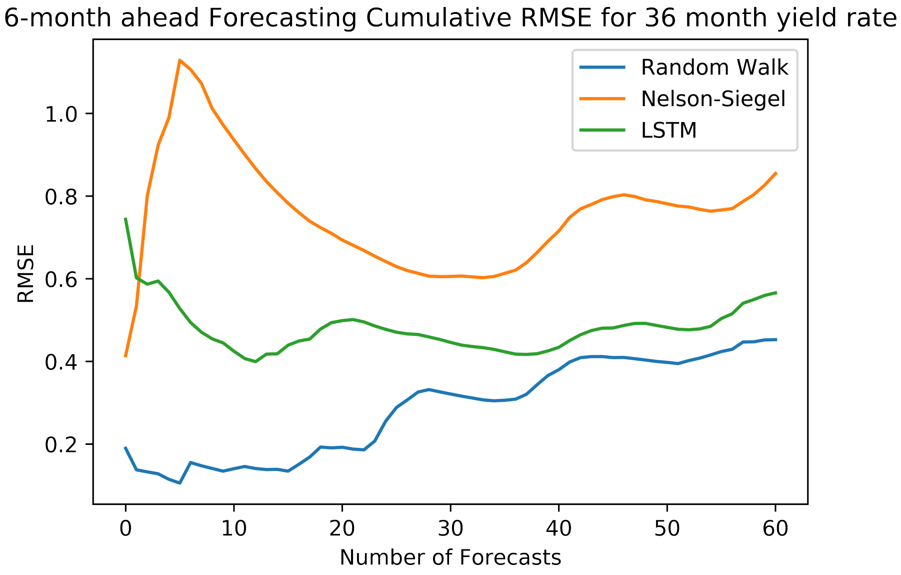


Figure 34: 6-month Ahead Forecast RMSEs for 36-month Maturity from 2015 to 2019

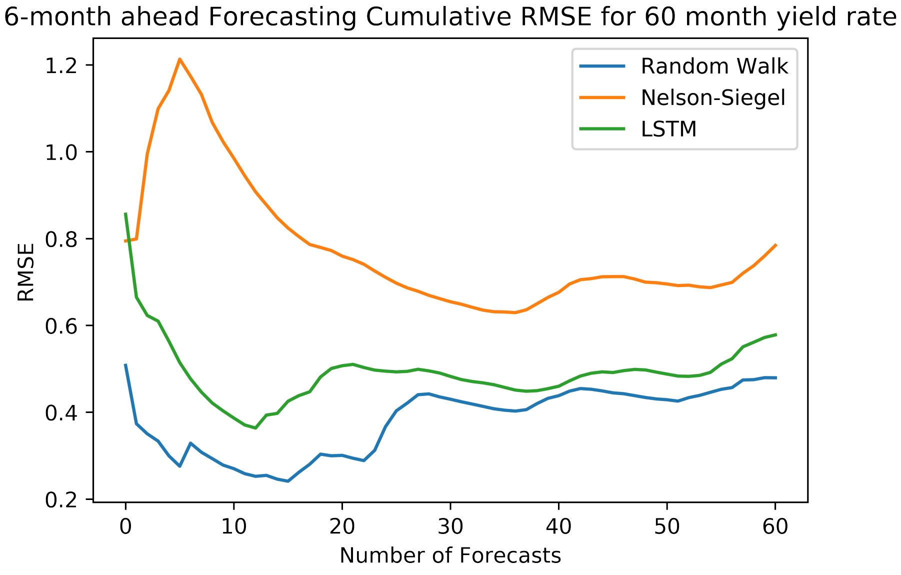


Figure 35: 6-month Ahead Forecast RMSEs for 60-month Maturity from 2015 to 2019

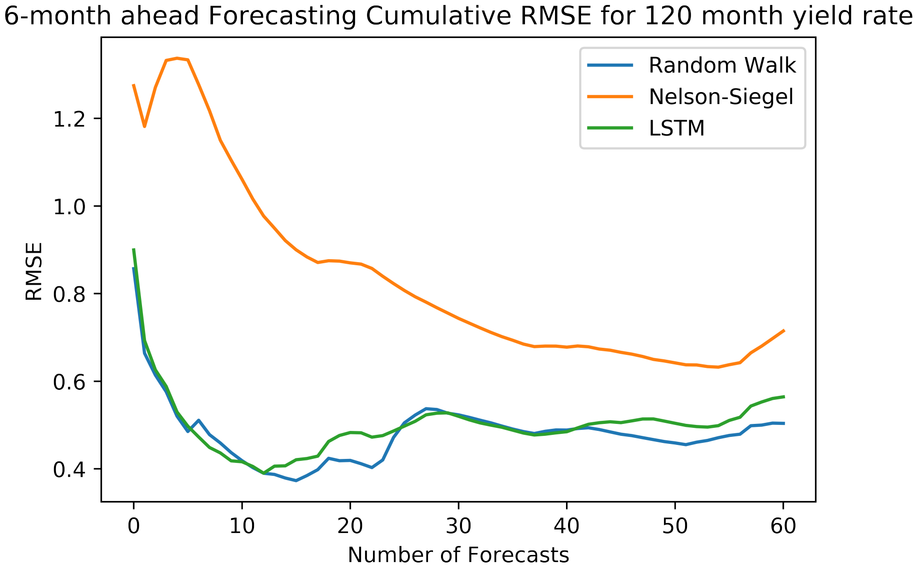


Figure 36: 6-month Ahead Forecast RMSEs for 120-month Maturity from 2015 to 2019

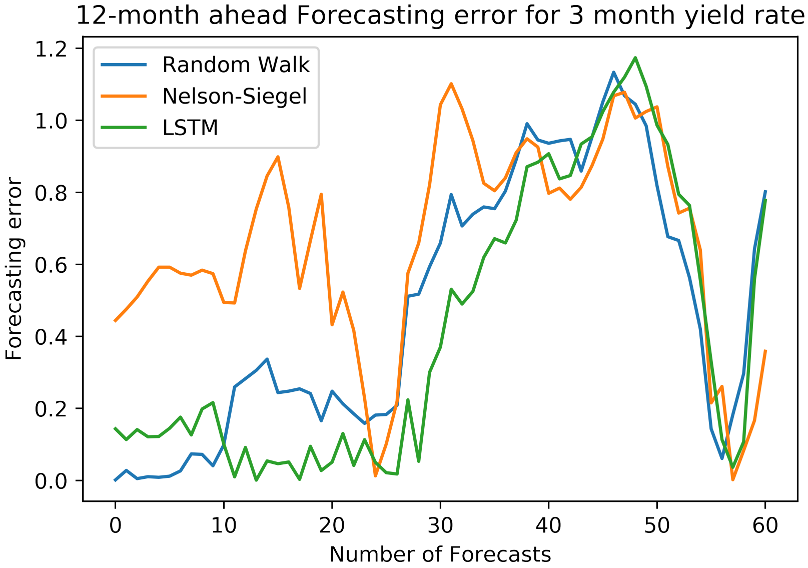


Figure 37: 12-month Ahead Forecast Errors of Each Period for 3-month Maturity from 2015 to 2019

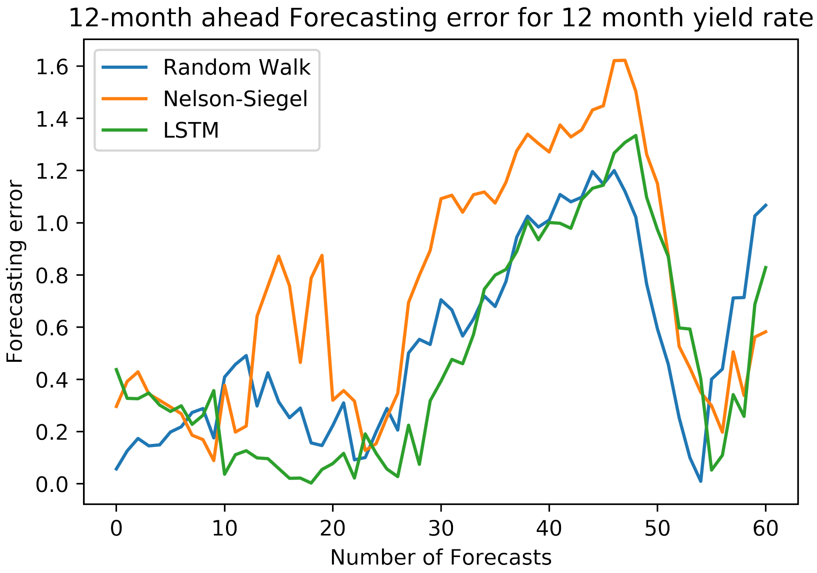


Figure 38: 12-month Ahead Forecast Errors of Each Period for 12-month Maturity from 2015 to 2019

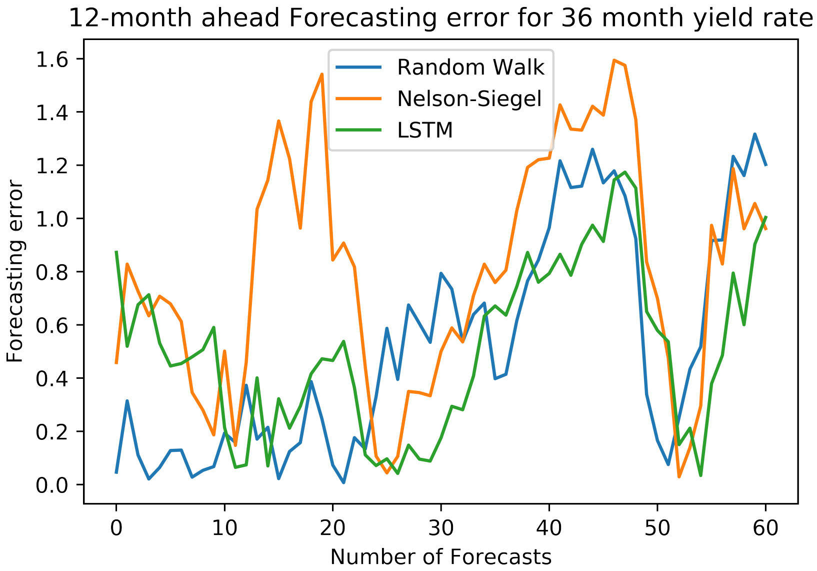


Figure 39: 12-month Ahead Forecast Errors of Each Period for 36-month Maturity from 2015 to 2019

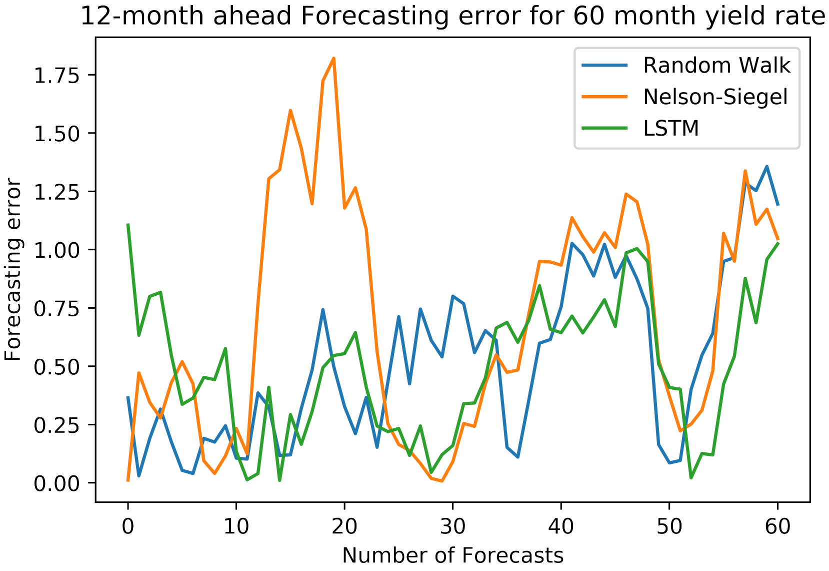


Figure 40: 12-month Ahead Forecast Errors of Each Period for 60-month Maturity from 2015 to 2019

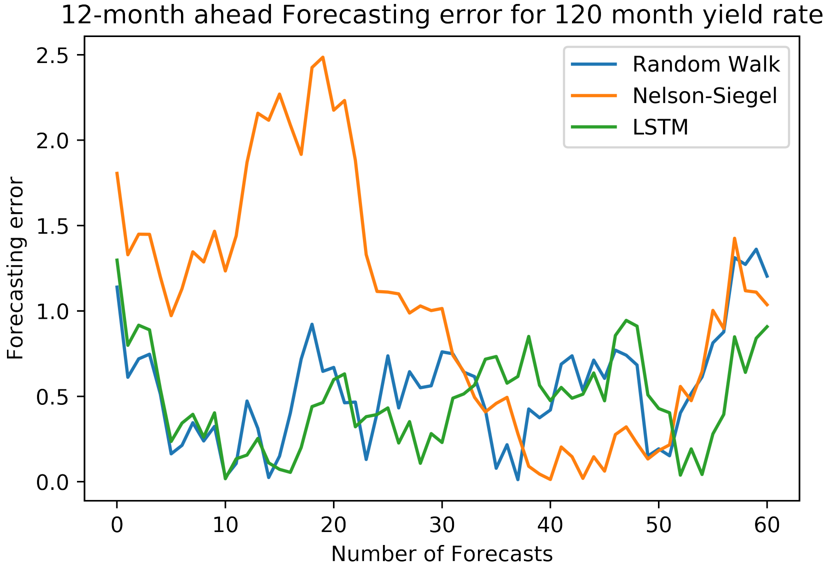


Figure 41: 12-month Ahead Forecast Errors of Each Period for 120-month Maturity from 2015 to 2019

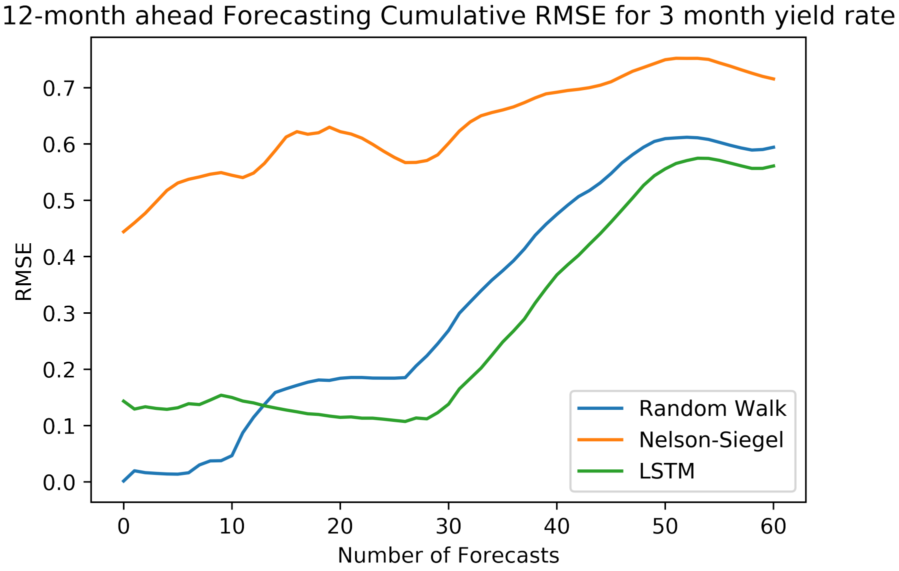


Figure 42: 12-month Ahead Forecast RMSEs for 3-month Maturity from 2015 to 2019

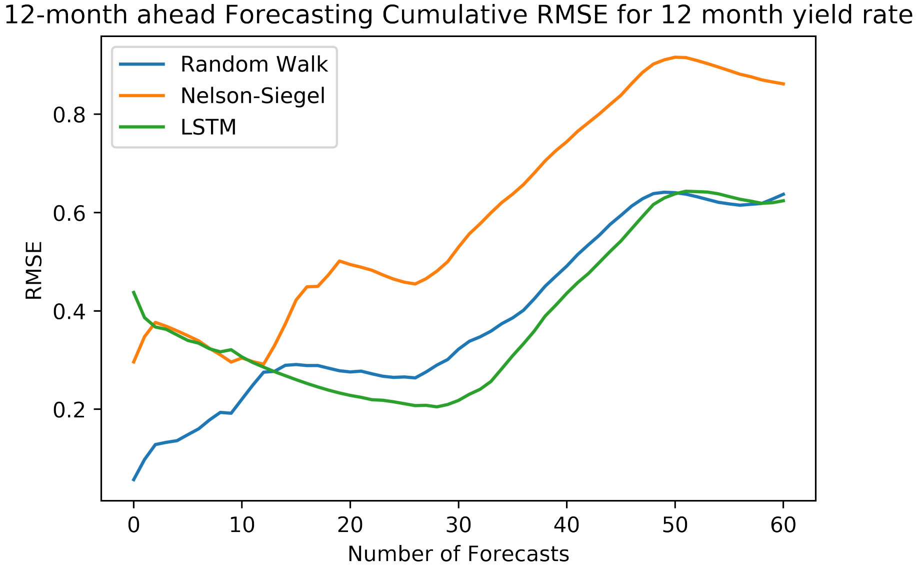


Figure 43: 12-month Ahead Forecast RMSEs for 12-month Maturity from 2015 to 2019

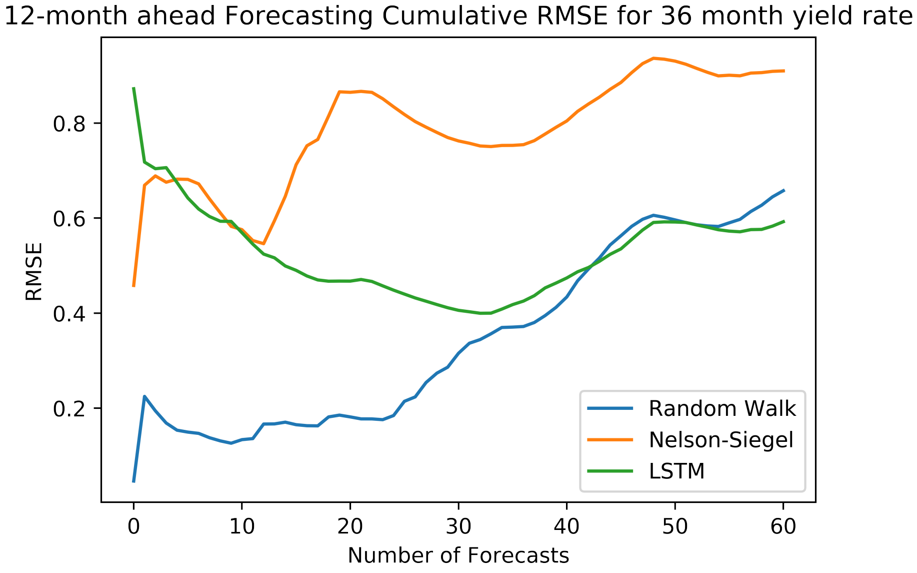


Figure 44: 12-month Ahead Forecast RMSEs for 36-month Maturity from 2015 to 2019

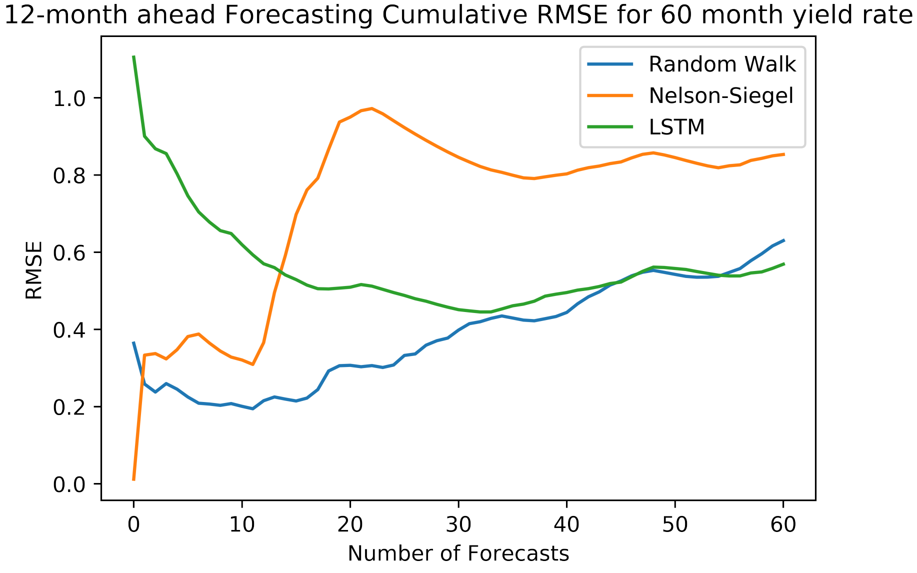


Figure 45: 12-month Ahead Forecast RMSEs for 60-month Maturity from 2015 to 2019

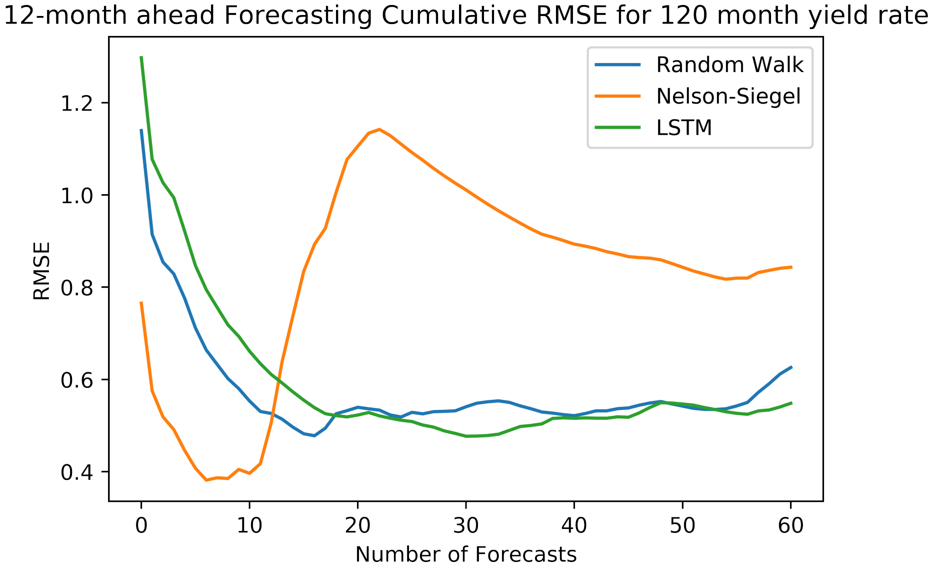


Figure 46: 12-month Ahead Forecast RMSEs for 120-month Maturity from 2015 to 2019

Bibliography

[1] Michael Anderson and Jeremy Magruder. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989, 2012.

[2] Rolph E Anderson. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *Journal of marketing research*, 10(1):38–44, 1973.

[3] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[4] Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

[5] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.

[6] Anol Bhattacherjee. Understanding information systems continuance: An expectation-confirmation model. *MIS quarterly*, pages 351–370, 2001.

[7] Daniele Bianchi, Matthias Büchner, and Andrea Tamoni. Bond risk premia with machine learning. *USC-INET Research Paper*, (19-11), 2019.

[8] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] Susan A Brown, Viswanath Venkatesh, and Sandeep Goyal. Expectation confirmation in technology use. *Information Systems Research*, 23(2):474–487, 2012.

[10] Benjamin Bushong and Tristan Gagnon-Bartsch. Reference dependence and attribution bias: Evidence from real-effort experiments. 2020.

[11] Meghan R Busse, Devin G Pope, Jaren C Pope, and Jorge Silva-Risso. The psychological effect of weather on car purchases. *The Quarterly Journal of Economics*, 130(1):371–414, 2015.

[12] Armand V Cardello and F Miles Sawyer. Effects of disconfirmed consumer expectations on food acceptability. *Journal of sensory studies*, 7(4):253–277, 1992.

[13] Richard N Cardozo, Ivan Ross, and William Rudelius. New product decisions by marketing executives: A computer-controlled experiment. *Journal of Marketing*, 36(1):10–16, 1972.

[14] J Merrill Carlsmith and Elliot Aronson. Some hedonic consequences of the confirmation and disconfirmation of expectances. *The Journal of Abnormal and Social Psychology*, 66(2):151, 1963.

[15] Edward C Chang, Albert Maydeu-Olivares, and Thomas J D'Zurilla. Optimism and pessimism as partially independent constructs: Relationship to positive and negative affectivity and psychological well-being. *Personality and individual Differences*, 23(3):433–440, 1997.

[16] Pei-Yu Chen, Yili Hong, and Ying Liu. The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science*, 64(10):4629–4647, 2018.

[17] Yiwei Chen. User-generated physician ratings—evidence from yelp. 2018.

[18] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[19] Jens HE Christensen, Francis X Diebold, and Glenn D Rudebusch. The affine arbitrage-free class of nelson–siegel term structure models. *Journal of Econometrics*, 164(1):4–20, 2011.

[20] Gilbert A Churchill Jr and Carol Surprenant. An investigation into the determinants of customer satisfaction. *Journal of marketing research*, 19(4):491–504, 1982.

[21] Thomas Cook and Aaron Smalter Hall. Macroeconomic indicator forecasting with deep neural networks. *Federal Reserve Bank of Kansas City, Research Working Paper*, (17-11), 2017.

[22] John C Cox, Jonathan E Ingersoll Jr, and Stephen A Ross. A theory of the term structure of interest rates. *Econometrica (pre-1986)*, 53(2):385, 1985.

[23] Jonathan Davis and Sara B Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50, 2017.

[24] Stefano DellaVigna. Psychology and economics: Evidence from the field. *Journal of Economic literature*, 47(2):315–72, 2009.

[25] Francis X Diebold and Canlin Li. Forecasting the term structure of government bond yields. *Journal of econometrics*, 130(2):337–364, 2006.

[26] Francis X Diebold and Robert S Mariano. Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144, 2002.

[27] Gregory R Duffee. Term premia and interest rate forecasts in affine models. *The Journal of Finance*, 57(1):405–443, 2002.

[28] Fdeloche. Long short-term memory. *Wikimedia Commons*, June 2017.

[29] Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4):93–106, 1962.

[30] Tristan Gagnon-Bartsch and Benjamin Bushong. Learning with misattribution of reference dependence. 2019.

[31] David Genesove and Christopher Mayer. Loss aversion and seller behavior: Evidence from the housing market. *The Quarterly Journal of Economics*, 116(4):1233–1260, 2001.

[32] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[33] Adam Goliński and Paolo Zaffaroni. Long memory affine term structure models. *Journal of econometrics*, 191(1):33–56, 2016.

[34] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[35] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021.

[36] Kareem Haggag, Devin G Pope, Kinsey B Bryant-Lees, and Maarten W Bos. Attribution bias in consumer choice. *The Review of Economic Studies*, 86(5):2136–2183, 2019.

[37] Jens Hainmueller, Jonathan Mummolo, and Yiqing Xu. How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis*, 27(2):163–192, 2019.

[38] Saram Han and Chris K Anderson. Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, 61(2):142–153, 2020.

[39] Yi-Chun Ho, Junjie Wu, and Yong Tan. Disconfirmation effect on online rating behavior: A structural model. *Information Systems Research*, 28(3):626–642, 2017.

[40] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[41] Tatiana A Homonoff. Can small incentives have large effects? the impact of taxes versus bonuses on disposable bag use. *American Economic Journal: Economic Policy*, 10(4):177–210, 2018.

[42] Carl I Hovland, OJ Harvey, and Muzafer Sherif. Assimilation and contrast effects in reactions to communication and attitude change. *The Journal of Abnormal and Social Psychology*, 55(2):244, 1957.

[43] Joop J Hox, Mirjam Moerbeek, and Rens Van de Schoot. *Multilevel analysis: Techniques and applications*. Routledge, 2017.

[44] Maggie Rong Hu and Adrian D Lee. Outshine to outbid: Weather-induced sentiment and the housing market. *Management Science*, 66(3):1440–1472, 2020.

[45] Nan Hu, Paul A Pavlou, and Jie Jennifer Zhang. On self-selection biases in online product reviews. *MIS Q.*, 41(2):449–471, 2017.

[46] Ying-Kai Huang. Attribution bias on online reputation systems. *Available at SSRN 3834091*, 2021.

[47] Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3):933–959, 2012.

[48] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, pages 263–291, 1979.

[49] Dmitrii Kochkov, Jamie A Smith, Ayya Alieva, Qing Wang, Michael P Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21), 2021.

[50] Botond Kőszegi and Matthew Rabin. A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165, 2006.

[51] Botond Kőszegi and Matthew Rabin. Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073, 2007.

[52] Botond Kőszegi and Matthew Rabin. Reference-dependent consumption plans. *American Economic Review*, 99(3):909–36, 2009.

[53] Anastasis Kratsios and Cody Hyndman. Deep learning in a generalized hjm-type framework through arbitrage-free regularization. *arXiv preprint arXiv:1710.05114*, 2017.

[54] Jonathan Lafky. Why do people rate? theory and evidence on online ratings. *Games and Economic Behavior*, 87:554–570, 2014.

[55] Nancy K Lankton, D Harrison McKnight, Ryan T Wright, and Jason Bennett Thatcher. Research note—using expectation disconfirmation theory and polynomial modeling to understand trust in technology. *Information Systems Research*, 27(1):197–213, 2016.

[56] George Leckie. Cross-classified multilevel models. *LEMMA VLE Module*, 12:1–60, 2013.

[57] David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

[58] Xinxin Li and Lorin M Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008.

[59] Mingxian Lin. Rnn pict. *Wikimedia Commons*, Dec 2018.

[60] Yan Liu and Jing Cynthia Wu. Reconstructing the yield curve. *Available at SSRN 3286785*, 2019.

[61] Long short-term memory. Long short-term memory — Wikipedia, the free encyclopedia, 2021. [Online; accessed 12-June-2021].

[62] Michael Luca. Reviews, reputation, and revenue: The case of yelp. com. 2016.

[63] John G Lynch Jr, Dipankar Chakravarti, and Anusree Mitra. Contrast effects in consumer judgments: Changes in mental representations or in the anchoring of rating scales? *Journal of Consumer Research*, 18(3):284–297, 1991.

[64] Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714, 2008.

[65] Vicki McKinney, Kanghyun Yoon, and Fatemeh "Mariam" Zahedi. The measurement of web-customer satisfaction: An expectation and disconfirmation approach. *Information systems research*, 13(3):296–315, 2002.

[66] Abu Hijleh Muhammad. Autoencoder. *Wikimedia Commons*, Jan 2017.

[67] Charles R Nelson and Andrew F Siegel. Parsimonious modeling of yield curves. *Journal of business*, pages 473–489, 1987.

[68] John R Nofsinger. Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3):144–160, 2005.

[69] Axel Ockenfels, Dirk Sliwka, and Peter Werner. Bonus payments and reference point violations. *Management Science*, 61(7):1496–1513, 2015.

[70] Shigehiro Oishi, Robert S Wyer Jr, and Stanley J Colcombe. Cultural variation in the use of current life satisfaction to predict the future. *Journal of personality and social psychology*, 78(3):434, 2000.

[71] Richard L Oliver. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of applied psychology*, 62(4):480, 1977.

[72] Richard L Oliver. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, 17(4):460–469, 1980.

[73] Jerry C Olson and Philip A Dover. Disconfirmation of consumer expectations through product trial. *Journal of Applied psychology*, 64(2):179, 1979.

[74] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[75] James L Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325, 1984.

[76] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.

[77] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[78] Christoph Schneider, Markus Weinmann, Peter NC Mohr, and Jan vom Brocke. When the stars shine too bright: The influence of multidimensional ratings on online consumer ratings. *Management Science*, 2020.

[79] Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.

[80] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

[81] Domonkos F Vamossy. Investor emotions and earnings announcements. *Journal of Behavioral and Experimental Finance*, 30:100474, 2021.

[82] Oldrich Vasicek. An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188, 1977.

[83] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[84] Wenche Wang, Fan Li, and Zelong Yi. Scores vs. stars: A regression discontinuity study of online consumer reviews. *Information & Management*, 56(3):418–428, 2019.

[85] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

[86] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural im-

age caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.