Evaluating Forecasting, Knowledge, and Visual Analytics

Yafeng Lu, Michael Steptoe, Verica Buchanan, Nancy Cooke, Ross Maciejewski
Arizona State University

Abstract

In this paper, we explore the intersection of knowledge and the forecasting accuracy of humans when supported by visual analytics. We have recruited 40 experts in machine learning and trained them in the use of a box office forecasting visual analytics system. Our goal was to explore the impact of visual analytics and knowledge in human-machine forecasting. This paper reports on how participants explore and reason with data and develop a forecast when provided with a predictive model of middling performance ($R^2 \approx .7$). We vary the knowledge base of the participants through training, compare the forecasts to the baseline model, and discuss performance in the context of previous work on algorithmic aversion and trust.

1 Introduction

Predictive visual analytics methods have been applied in a variety of domains ranging from healthcare [1], intelligence analysis [4], and emergency crisis management [35], and a great deal of research in the visual analytics community has focused on explaining predictive models [5, 27, 31, 36, 42, 49] and enabling interactive model steering [6, 23, 24, 30, 44, 46]. A variety of research methods advocate adding human-machine interactions not only to improve model understanding but also to enable knowledge injection into the system [23, 26]. This intersection of human-machine analysis is seen as a critical stage in the predictive visual analytics pipeline [32]. Yet, this potential for adding knowledge comes with increased risk. Allowing humans to assert their knowledge constructs into the prediction process may inherently bias the model itself [14, 41] as humans, while a wealth of contextual information, are biased in their own thought and knowledge [7, 17]. If human input is too closely tied to the prediction then the prediction may become biased in its assumptions and may become less accurate.

Such concerns are buoyed by research in the decision science field that has shown that in forecasting tasks, machine predictions consistently outperform human forecasters [8,9,19,43]. In fact, work by Akes, Dawes, and Christensen [2] found that domain expertise diminished people's reliance on algorithmic forecasts which led to a worse performance. Studies have also shown that humans develop an algorithm aversion in forecasting tasks [11, 20]. Specifically, humans quickly lose confidence in algorithmic forecasts after seeing algorithmic mistakes [50]. Given that the underlying goal of many predictive visual analytics methods is to inject knowledge into the analysis and point out potential algorithmic errors to the analyst for updating and correction, these goals may be at odds with human behavior. As such, visual analytics could potentially contribute to algorithmic aversion during forecasting tasks and lead to reduced performance. Conversely, studies report that forecasters may desire to adjust algorithmic outputs to gain a sense of ownership of the forecasts due to a lack of trust in statistical models [12], and the goal of many predictive visual analytics methods is to help develop trust.

Given the conflicting demands of model accuracy, comprehensibility and trustworthiness, the question of how much human knowledge and interaction is needed or warranted in relation to the model becomes a critical question for predictive visual analytics. Are humans able to accurately make predictions or outperform models with the aid of visual analytics, and what is the impact of knowledge in a visual analytics forecasting process? We seek to better understand the effects of knowledge and prediction accuracy when using visual analytics support for human-machine forecasts. Inspired by work from the 2013 VAST Box Office Challenge [13, 28, 34], the goal of this study is to explore forecasting in a visual analytics setting. Based on studies from management, economic sciences and psychology, algorithmic forecasts commonly outperform humans. However, the results in the VAST Challenge [35] indicated that by leveraging visual analytics methods, subjects were able to improve forecasts from models, and a study exploring managerial intervention in sales forecasting also found an increase in the overall prediction accuracy [37] was achieved through human intervention. As such, we hypothesized that with a middle level performance model, similar to the performance of models used in related forecasting studies [8, 9, 19, 43], a visual analytics forecasting process may serve as a more optimal human-machine configuration.

A controlled study was conducted to test the hypothesis that a predictive visual analytics framework supporting human-machine prediction will outperform the computational solution given by a model at a middling level of accuracy. Given that visual analytics suggests the integration of knowledge as part of the process, participants were only chosen from a pool having relevant knowledge in forecasting to allow for a baseline control. Knowledge related to the subject matter was then varied across participants to further explore the intersection of knowledge, forecasting, and visual analytics. In this paper, the term knowledge is explicitly referring to the concept of tacit knowledge [15, 48].

The visual analytics system used in this study is modified from previous work [3, 33], and a prediction model with a 70% goodness-of-fit was developed. Participants were asked to predict the opening weekend gross of 4 movies in the same genre with similar levels of popularity using our system. This paper presents the results from the study and serves as a starting point to discuss open challenges of the role of visualization in human-machine forecasting.

2 Background

The field of human factors has long been interested in the relationship of humans and technology, and the effect of biases on human decision-making has been studied throughout multiple contexts. In particular, confirmation bias (seeking out information to confirm decisions [7, 38]), overconfidence bias (being too confident in abilities which leads to taking risks [25]), and anchoring (over-reliance on first piece of information found [17]) are specific human biases that are known to affect decision-making. With respect to predictive analytics, each of these biases may impact how humans utilize different analytical tools and predictive models.

A great deal of human factors work has also focused on how humans trust machinery, specifically automation and autonomy [39]. Recent work by Hoff and Bashir [22] identified that human's trust in automation is highly dynamic and dependent on a multitude of factors. Those factors can be distilled into three main areas of trust: dispositional trust (dependent on culture, age, gender, personality), situational trust (type of system, system complexity, task, etc.), and learned trust (past or current experience with systems).

With these uncertain human factors, researchers have also studied the reasons why people prefer to use human predictions over automatic models in many scenarios [14, 29]. One reason is that humans seem to have an inherent distrust of algorithmic models, and examples of this distrust are found in various fields including organizational planning [16], hiring [21], and clinical predictions [47]. Here, people rely on their judgment and intuition much more than prediction algorithms, although investigations show that the prediction performance could be improved if they relied more on computational models. Other studies have also indicated that lack of trust stems from the limitation of automatic techniques and challenges of model explainability [18, 40].

This algorithm aversion phenomenon is further discussed by Dietvorst, Simmons and Massey [11] where studies indicate that people are less likely to use forecasts from an algorithm after seeing it perform and learning that it is imperfect, even if they also see that it outperforms the human forecaster. A related study [12] found that people are much more willing to use forecasts from an imperfect algorithm when they can retain a slight amount of control over the algorithm's forecasts. This study found that letting people adjust an imperfect algorithm's forecasts would increase both their chances of using the algorithm and their satisfaction with the results. However, the authors also found that participants in the study often worsened the algorithm's forecasts when given the ability to adjust them. Dietvorst [10] studied the decision process that leads people to rely on human predictions instead of algorithmic predictions. In this study, it was found that the prediction method used depends on (1) the status quo prediction method which is a default choice, and (2) whether an alternative method can meet people's counter-normative reference points. Given these results, it is clear that more studies are needed to provide guidelines for methodologies and designs when supporting forecasting with visual analytics.

3 Experimental Design

The goal of this experiment is to identify the impacts of tacit knowledge in a predictive visual analytics setting. In this experiment, we control for tacit knowledge by exclusively selecting participants who have knowledge in forecasting. Graduate students in the business school, industrial engineering and computer science with expertise in data analysis and forecasting were recruited. All participants had strong knowledge of linear regression models, and had applied these in their own data analyses. This means that all participants had a baseline domain expertise relevant to the type of model used in the experiment (i.e., linear regression model).

After controlling our subject pool to ensure that our participants had similar tacit knowledge with respect to forecasting, we then controlled for domain knowledge by providing half of the participants with additional training on movies in order to emulate specific subject (movie) domain knowledge (as compared to the broad forecasting domain knowledge that all participants have). Participants were then required to use a web-based visual analytics system to predict the opening weekend gross of a movie. A baseline computational model is given as a reference, and a set of visual analytics tools on the movie's meta-data and twitter data is provided. We then collected system interactions and forecasts for four different movies in order to explore the following research questions:

- Can a participant (with visual analytics support) develop more accurate forecasts than a model of middling level accuracy?
- Can a participant with extra knowledge develop more accurate forecasts than a participant without such knowledge given the same visual analytics environment?

Here, we note that it is almost impossible to emulate an expert with years of experience in a user study. However, access to a large sample size of domain experts is difficult to obtain (and impossible for some domains). What we feel is interesting about this study is that (to our knowledge), this is the first attempt to control for the effects of knowledge in a human-machine forecasting setting. Compared to previous studies, which were conducted on Amazon Turk [10] or on business school students [11], we focused on controlling for knowledge in predictive analytics. Previous work by Dietvorst [10] did not control for expertise, and Dietvorst et al. [11] assumed that business school students would have some intuition about MBA admissions (however, there was no guarantee of expertise). By selecting participants who have domain expertise in forecasting and augmenting participants' knowledge of predictive analytics with knowledge that will be directly relevant to the prediction, participants with the augmented knowledge should have an advantage in the prediction process.

By splitting our predictive analysis domain experts into two groups, we can explore the potential impact of extra domain knowledge. This is not to say that some participants are not domain experts. All participants in this study are experts in predictive analytics; however, recruiting box office prediction experts at the size and scale needed for a formal user study is not realistic. Many domains explored in predictive analytics often have an extremely limited number of domain experts (and box office prediction is no different). However, the general population has some inherent knowledge of movies, and we have further selected participants based on direct knowledge of predictive analytics. By recruiting participants that have domain knowledge in prediction and then adding specific knowledge (that was designed to be useful), we can begin exploring the differences between two such groups. Although there are limitations in this proposed methodology, this study provides (to our knowledge) the first attempt to compare the impact of different types of knowledge in forecasting tasks.

3.1 Hypotheses

Our experiment examines the role of humans (compared to automatic models and/or pre-defined approaches) and the effect of participant knowledge in predictive visual analytics. In this study, our hypotheses were tested by comparing the performance between participants (all of which are experts in predictive analytics) where some participants are provided with specific domain knowledge.

3.1.1 Prediction Performance

Visual analytics use cases have shown that interactive analytics can contribute to intelligence analysis by integrating domain knowledge through feature selection and subsequent model adjustments. Hence, participants possess the ability to analyze real world prediction problems more comprehensively than relying solely on a model. The impact of the participant's knowledge should be reflected in the prediction performance and having domain knowledge is hypothesized to contribute positively to the prediction accuracy.

Hypothesis 1: Participants will make more accurate predictions than purely algorithmic models when using visual analytics.

Numerous studies indicate that participants' predictions are generally worse than model predictions, and at the same time, others cite the opposite. Additionally, some research has shown that participants' confidence and satisfaction improve after being allowed to make changes based on a model's prediction [12]; however, prediction outcomes do not tend to improve. As a result, participants' contributions to prediction accuracy may be limited. Our goal is to further explore how participants' knowledge plays into the predictive analytics task in a visual analytics environment.

Hypothesis 2: Participants with more domain knowledge will make more accurate predictions than participants with less domain knowledge.

Prediction performance has been claimed to be dependent on the integration of participants' domain knowledge. We hypothesized that participants with more domain knowledge will predict more accurately than participants with less knowledge.

3.1.2 Algorithm Aversion with Domain Knowledge

In the previous research, Akes, Dawes, and Christensen [2] found that domain expertise diminished people's reliance on algorithmic forecasts which led to a worse performance. It is reasonable to expect that a participant could refer to the model's prediction as one important factor in making their own predictions. However, the anchoring to the computational model prediction is hypothesized to be less when the participant has more domain knowledge because they are more likely to notice errors in the model and prone to be more confident in making adjustments. As such, we expect that participants with more domain knowledge will adjust the forecasts more than participants with less domain knowledge.

Hypothesis 3: Participants with more subject knowledge will apply greater adjustments to the forecast than participants with less subject knowledge.

3.2 Experimental Design Factors

In order to test our hypotheses, we developed a predictive model with an approximately 70% goodness-of-fit. Additionally, we designed the experiment with two experimental groups with each group having expertise in data analysis but one set of training was designed to simulate movie expertise. The task was to utilize a visual analytics system to make predictions of movies' opening weekend gross.

3.2.1 Dataset

Our box office prediction task uses movie meta-data from the Internet Movie Database (IMDB) and social media posts from Twitter. The meta-data includes release date, genre, MPAA rating, and estimated budget. For social media data, following the data collection strategy of the 2013 VAST Box Office Challenge [34], we have collected movie related tweets for 388 movies in the United States released from January 2013 to April 2017. Tweets are crawled using the Twitter Streaming API [45] by searching the hashtag extracted from each movie's official Twitter account. In this study, we use tweets posted two weeks prior to each release date.

3.2.2 Baseline Model in the Experiment

Linear regression analysis was applied on a variety of factors (Table 1) from both twitter data and movie meta-data to fit the opening weekend gross. Our linear regression model uses a square root data transformation on the response for a normal residual distribution.

Table 1: Variables used in the Experimental Visual Interface

Variable	Description
Gross	3-day Opening Weekend Gross
Budget	Approximate movie budget in M dollars
Screen	The number of theaters a movie is released in
TBD	The average daily number of Tweets over the 2 weeks prior to release

Baseline Model:

$$gross_{sr} = \beta_0 + \beta_1 Budget + \beta_2 TBD + \beta_3 Screen + \varepsilon.$$

This model has $R^2 = 71.71\%, R^2_{adj} = 71.49\%, R^2_{pred} = 70.46\%$.

3.2.3 Movie Knowledge Training

As previously stated, our experimental design was to randomly assign participants into two groups, the Data Group and the Movie Group. Both groups have knowledge in forecasting; however, the Movie Group participants were provided extra knowledge during training to simulate domain expertise . Training information was developed to provide insights into movie performance that may not be obvious to experts with only a data analysis background. As part of the Movie Group condition, participants were presented a 19-page training document containing knowledge designed to be useful in their predictions, such as how release time, etc. could affect revenue.

To validate that they gained knowledge, participants in this group were given a quiz about the training material. The quiz had 6 multiple choice questions. 17 participants scored 6/6 and 3 scored 5/6. Unknown to the participants, the knowledge provided was specific to the movies they would encounter during the analysis and was designed to give insight into if these movies over-performed or under-performed with respect to the baseline model. For example, participants were told that movies that engage in counter-programming (such as action movies on Valentine's Day weekend) over-perform on box-office returns. While this knowledge may not always be true, for the movies in our forecasting challenges, this concept is accurate. Thus, we simulate additional knowledge in one set of participants (again, all participants have expertise in forecasting) and compare this to participants who have domain knowledge in data analysis but have no extra movie domain specific information. In the control condition, participants receive training that is equally long but provides no relevant domain knowledge.

3.2.4 Other Factors

Participant bias: To mitigate participant bias, we narrowed down our subjects to be full-time undergraduate and graduate students with data analytics skills. Specifically, each subject was required to have minimally completed a course in machine learning, data mining, or statisical forecasting. In addition, each participant estimated four movies' opening weekend gross to reduce randomness.

Movie bias: A movie, as the object of this prediction task, might also impact the performance and variance of different participants' predictions. In order to mitigate such movie bias, we selected particularly well-known movies based on a high Tweet count and all 4 movies are from the same genre (Action).

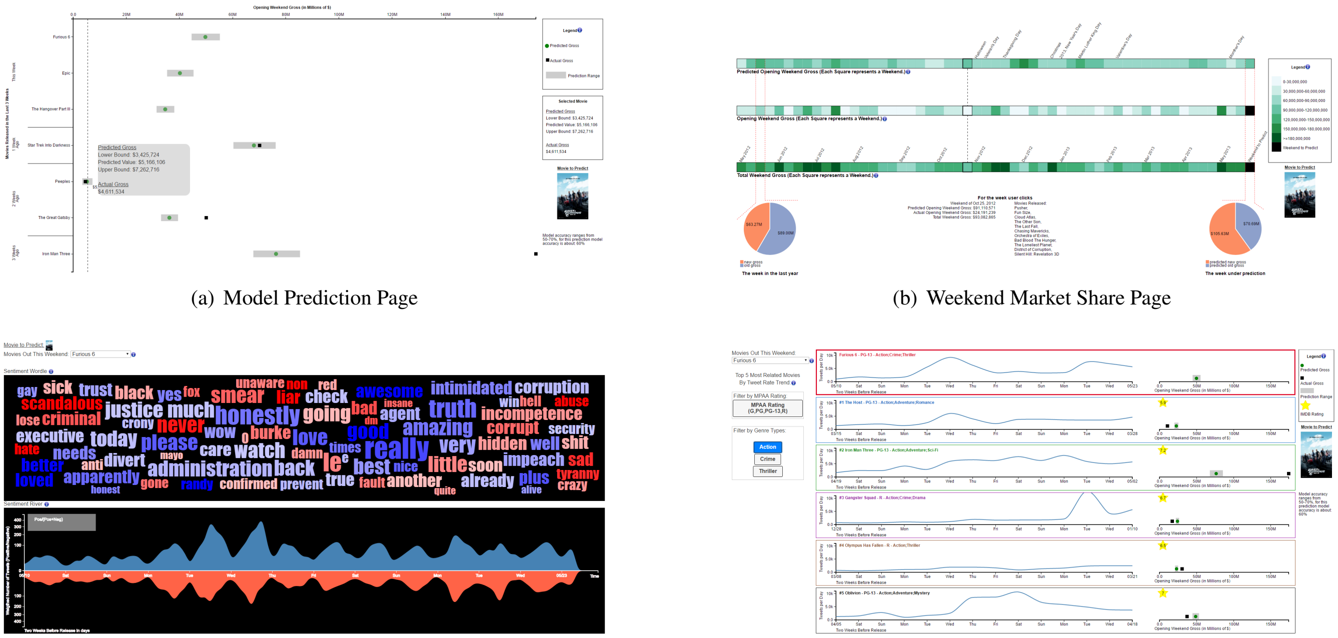
Order bias: A major concern for repeated prediction tasks is that participants will become familiar with the system and task, or participants may become fatigued as more time is spent on these tasks. This can lead to performance variation simply due to the order the movies were presented. To avoid such bias, we randomized the order in which the movies were presented.

3.3 Visual Analytics Interface Design

The experiment reported in this article utilizes a previously developed visual analytics framework (where the visual designs were comparable to other teams participating in the contest) for box office analysis and prediction [3] that has been modified to: (1) facilitate the recording and implementation of the experiment, and; (2) provide a convenient way for analysts to explore and analyze the data. The visual analytics framework used has been extensively tested in various forms for usability, and survey results from this study (and others) indicates that the system is easy to use.

This visual analytics system consists of six visual components: Homepage, Model Prediction, Weekend Market Share, Sentiment Analysis, Movie Similarity, and Make Prediction. The system uses pre-processing to extract useful information from large-scale, noisy, and unstructured Twitter data. These data are then integrated into the visual analytics system for easy exploration and knowledge extraction. We have used numerical and nominal features shown in Table 1 that are extracted from IMDB and Twitter.

The *Homepage* shows basic information for the weekend under prediction and a system tutorial. It contains the date of the weekend and a brief introduction to the released movies on that weekend. The *Make Prediction* page is where participants submit final predictions. This page lists the numerical prediction result of the baseline model and the weekend market share model (which is used in the Weekend


Opening Weekend Gross (in Millions of $)
Legend
Predicted Gross
Actual Gross
Prediction Range
Selected Movie
Movie to Predict
Furious 6
Epic
The Hangover Part III
Star Trek Into Darkness
Peeples
The Great Gatsby
Iron Man Three
Predicted Opening Weekend Gross (Each Square represents a Weekend.)
Opening Weekend Gross (Each Square represents a Weekend.)
Total Weekend Gross (Each Square represents a Weekend.)
The week in the last year
The week under prediction
Sentiment Wordle
Sentiment River
Movies Out This Weekend
Top 5 Most Related Movies By Tweet Rate Trend
Filter by MPAA Rating
Filter by Genre Types
Action
Crime
Thriller
Two Weeks Before Release

(a) Model Prediction Page

(b) Weekend Market Share Page

(c) Sentiment Analysis Page

(d) Movie Similarity Page

Figure 1: The four main visual components in the experiment for data exploration and predictive analytics.

Market Share page). These two predictions can be referred to by the participant while analyzing the data and making a decision without being restricted by any model predictions. The other pages contain visualizations and the participants can navigate between them freely.

The *Model Prediction* page (Figure 1(a)) shows the baseline model's results and performance for movies released in the current week and previous three weeks. This page orders the movies by their release dates along the y-axis with current week on top and separate movies by each week using solid lines, and it plots the model's prediction value along the x-axis. The green circle and surrounding gray bar show the model's prediction value and its 95% confidence interval, whereas the black squares indicate the actual opening weekend gross for previous movies. Mousing over the plot shows a dashed line referring to the gross axis. To look at the exact values of a movie the participant can click on the green circle to open a pop-up context window.

The *Weekend Market Share* page (Figure 1(b)) visualizes the result from a temporal model of weekend market prediction for the participants to identify seasonal patterns. This page has three horizontal bars where each one consists of 55 squares covering a whole year's weekends prior to the weekend under prediction. The revenue of each weekend is shown by the color of its square according to a sequential color scale where the light color means low revenue and the dark color means high revenue. The three horizontal bars correspond to the temporal model's prediction, the real value of the sum of all newly released movies, and the real total weekend gross of all currently playing movies. Mousing over these squares will line up the weekends from these three bars and clicking on the square can highlight this weekend and display the details of the released movies and the revenue of that weekend. Holidays are marked on top for special day highlights. The exact values of the temporal model's prediction are visualized as pie charts for the current weekend and the corresponding weekend in the last year.

The *Sentiment Analysis* page (Figure 1(c)) consists of a sentiment wordle and a sentiment river plot [35]. Here, participants are provided with an overview of the public's expectation of a movie. Positive/negative sentiment is shown as blue/red.

The visual analytics environment also supports the comparison between movies in the *Movie Similarity* page (Figure 1(d)). Similar movies can be filtered by selecting the movie's MPAA rating and its genre(s). Once a metric is selected, the movies are first filtered by these metrics and the five most similar movies ordered by tweet volume trend are displayed. The left side of this page lists the options of filtering for similar movies and the right side uses small multiple views to show the five most similar movies and the current movie under prediction (the top one). The view of each movie contains a line chart of the tweet volume trend and the prediction.

4 Experiment

Forty participants were recruited and randomly assigned into two groups (Data Group and Movie Group) with 20 participants in each. The Data Group received training on how to write a movie script (but not about box office prediction) and the Movie Group received training pertaining to box office forecasting. Each group had 5 female participants and 15 male participants. The Data Group had 1 undergraduate and 19 graduate students and the Movie Group had 2 undergraduates and 18 graduate students. The average age was 26.3 years old, ranging from 19 to 31. Each participant participated in a training session, a practice prediction session, and four experimental prediction sessions. Their workload was evaluated twice using the NASA Task Load Index (NASA TLX) measure; once after the practice session and once after the last prediction was cast. The participants also completed a demographic questionnaire about their background, knowledge of predictive analytics, computer usage, and familiarity with the presented movies. Each session took approximately 2.5 hours.

4.1 Training

Each participant was trained at the start of the experiment. Training consisted of two parts, a domain knowledge training and a system usage training. Both training materials were presented using PowerPoint slides. After the training, the researcher loaded the visual analytics interface and participants were instructed to access the system "Tutorial" page in order to ensure that they knew where it

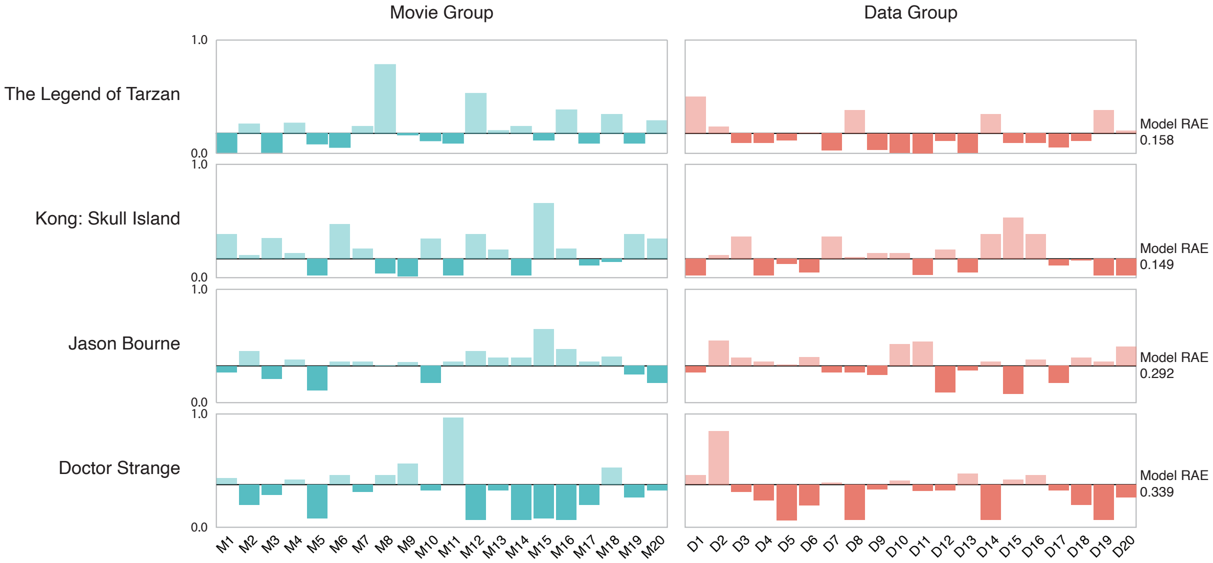


Figure 2: The RAE of each prediction organized by movie and data group. Each bar chart shows the RAE of one group participants' estimates for one movie, and the horizontal black line indicates the model's error. If the bar is below the line, the participant's forecast is better than the model (darker color), and vice versa.

was located. The researcher also pointed out the timer which was set for each session (15 minutes). The training PowerPoint and the "calculator" were also opened for participants' reference. Participants were also provided with scratch paper (for notes/calculations) and a prediction sheet to write down their final prediction.

4.2 Predictions

Exploration and predictions consisted of a practice prediction and four real predictions. The participants started with a 15-minute practice session, during which participants were given the movie *John Wick: Chapter Two* to explore. An embedded timer is shown in the system and participants were encouraged to finalize their prediction before 15 minutes were up. However, the system was designed so that participants could take longer to cast their prediction if they needed to. We allowed participants to take longer because our primary focus of this study was to understand the human aspect during the predictive visual analytics task. Participants were also encouraged to ask questions to ensure comprehension of the task and the interface. After the practice session, the participants were given a 10-minute break.

After the break, the participants completed four predictions. For each prediction, participants were encouraged to submit their answer within 15 minutes. The following movies were presented to the participants in a random order: *The Legend of Tarzan*, *Jason Bourne*, *Doctor Strange*, and *Kong: Skull Island*. After the experiment, the participants' written prediction sheets were collected. The system records the prediction and interactions with each interface feature.

4.3 Questionnaires

To evaluate the participants' mental workload, a NASA TLX was administered twice during the experiment, the first time after the practice session and the second time after the last prediction. It was expected that participants with domain knowledge (i.e., the movie group) experience less stress and workload and have more confidence in completing the tasks. A demographics questionnaire was also administered at the end of the experiment to evaluate the participants' background (age, gender, education), domain knowledge (movie familiarity, frequency of "going to the movies"), social media usage (frequency), and knowledge of predictive analytics (familiarity with mathematical models). Finally, participants were also asked a free-response question aimed to assess how they analyzed the data, and elicit strategies used to analyze the data.

5 Results

Our analysis uses the predictions made on a total of 160 predictions for four movies (which excludes the practice session). The demographics questionnaire and the NASA TLX evaluations were used to gain a deeper understanding of how participants worked through the analysis.

5.1 Prediction Performance

To test hypothesis 1, the participants' predictions were first compared to the model predictions. To test hypothesis 2, the predictions in the Data Group were compared to the predictions in the Movie Group. The relative absolute error (RAE), which is the percentage error deviating from the real value, is used to measure the accuracy.

$$RAE = \frac{|Prediction - RealValue|}{RealValue} \tag{1}$$

Figure 2 displays the RAE for each prediction. To test our first prediction performance hypothesis, a two-sample t-test with equal mean as the null hypothesis on the participants' RAE and the model's RAE was applied. The statistic result is given in Table 2. The result shows no significant difference with a p-value = 0.965. **Therefore, the first hypothesis that participants will make more accurate predictions than purely algorithmic models when using visual analytics is not supported.** While the difference between the mean RAE is only 0.0029, the standard deviation in participant predictions is as large as 0.162 compared to a 0.237 mean value. This indicates a large variance in participants' performance, as seen in Figure 2.

Next, we performed a two-sample t-test between the predictions in each group and the predictions given by the model and found no significant difference between their means (Table 2). **Thus, the second hypothesis that participants with more domain knowledge will make more accurate predictions than participants with less domain knowledge is not supported**.

What is interesting is that results from testing Hypothesis 1 does not completely align with previous work [8, 9, 19, 43] which found significant differences between machine predictions and human predictions (previous studies used model accuracies in the range of 70%

Table 2: Results of four two-sample t-tests for participant prediction performance compared to the model performance and between the two participant groups. RAEs are used as the samples and equal mean is the null hypothesis.

Sample	N	Mean	StDev	P-Value
Participant RAE	160	0.238	0.162	0.965
Model RAE	4	0.235	0.095	
Movie Group RAE	80	0.253	0.167	0.740
Model RAE	4	0.235	0.095	
Data Group RAE	80	0.222	0.156	0.817
Model RAE	4	0.235	0.095	
Data Group RAE	80	0.222	0.156	0.220
Movie Group RAE	80	0.253	0.167	

– 83%). Furthermore, results from testing Hypothesis 2 does not fully align with previous work [2] where domain expertise led to worse performance in forecasting. In our study, while we did not find a significant difference between the participants' predictions and model predictions, the fact that participants were not significantly worse may indicate that other factors bolstered participants' responses. Previous findings [12] found that people are much more willing to use forecasts from an imperfect algorithm when they can retain a slight amount of control over the algorithm's forecasts. Given that visual analytics allows participants to explore the data and model, our findings hint towards improved explainability and trust being developed as part of the visual analytics process, resulting in (potentially) more trust in the model (which may be why the participants' forecasts were not significantly different than the machine forecasts). Further studies are needed to confirm/reject these relationships.

5.2 Algorithm Aversion of Domain Knowledge

Our third hypothesis directly explores trust in the model through the proxy of a participant's adjustment. The computational model used in our experiment was approximately 70% accurate. On average, participants made a 19.2% adjustment from the model's prediction to their own prediction. The largest change was 80.4% made by participant M11 for Doctor Strange. There were only 5 predictions out of the 160 estimates that were adjusted by less than 2%, and 22 predictions were adjusted by less than 5%.

Work by Akes, Dawes, and Christensen [2] found that domain expertise diminished people's reliance on algorithmic forecasts which led to a worse performance. While our results indicate that the participants' forecasts are not significantly different than the model (meaning that domain expertise did not lead to a significant difference in performance), we also tested to see if the amount of domain knowledge (Movie Group vs. Data Group) would result in more adjustments to the forecast, indicating more algorithm aversion. To explore this, the correlation between the participant's prediction and the model prediction was analyzed. The Pearson correlation between the Movie Group predictions and model predictions is 0.578 with participants adjusting each forecast by an average of 20.25%. The Pearson correlation between the Data Group predictions and the model predictions is 0.664 with participants adjusting each forecast by an average difference of 17.87%. There was no statistical significance between the two groups. **Therefore, our third hypothesis that participants with more subject knowledge will apply greater adjustments to the forecast than participants with less subject knowledge is not supported**. As such, our findings do not fully support the results of previous work [2] where domain experts showed diminished reliance on the algorithmic forecasts.

5.3 Demographics Questionnaire Analysis

To further explore domain expertise, we also tested whether participants significantly differed among each other in terms of their prediction performance and familiarity of movies based on their own self-assessment measure. For each estimate, we had asked our participants if they were familiar with the movie; from the total 160 estimates, 86 marked that they were familiar and 74 marked that they were not. We applied a two-sample t-test on participant prediction RAE between familiar and unfamiliar and we found no significant difference (p-value = 0.200) between those self-reporting as having movie knowledge and those self-reporting as not.

5.4 NASA TLX Analysis

Finally, we compared and evaluated the mental effort and task demands of the forecasting process using the NASA TLX workload measure. We found that mental effort and task demands significantly differed across the two groups (Movie and Data Group). There was a significant difference between Data Group participants (M=7.00, SD=1.83) and Movie Group (M=5.45, SD=1.89) participants, t(78)=-3.73, SE=.42, p=.00, d=.83 in their mental effort rating. Overall, Data Group individuals found that predicting movie earning amounts was more taxing than the Movie Group individuals. There was also a significant difference between Data Group participants (M=6.68, SD=1.46) and Movie Group (M=7.4, SD=1.46) participants, t(78)=2.22, SE=.33, p=.03, d=.50 in their accomplishment rating. Here, Data Group individuals viewed themselves as less successful at predicting movie earning amounts than the Movie Group individuals. Lastly, there was a significant difference between Data Group participants (M=6.63, SD=2.63) and Movie Group (M=7.83, SD=2.04) participants, t(73.43)=2.28, SE=.33, p=.03, d=.51 in their emotional state rating. Data Group individuals felt more stressed than the Movie Group individuals during the prediction tasks. This may indicate that increased knowledge is leading to over-confidence bias resulting in higher (on average) deviations from the model which would be consistent with previous research [2] that found that domain expertise diminished reliance on algorithmic forecasts.

6 Conclusions and Future Work

Work in the visual analytics community has demonstrated that experts could benefit from tools that support the integration of domain knowledge with interactive visual exploration. In terms of predictive analytics, visual analytic techniques have also been used to help improve the comprehension of data, model, and prediction results. Visual analytics techniques that enable participants to interact within the predictive modeling process have also reported benefits. However, the social and management science communities have deliberated for decades as to whether or not humans can improve algorithmic prediction. Unfortunately, the literature does not provide a clear answer, and the impacts of visual analytics in the forecasting process have yet to be fully explored.

This study represents an initial step in exploring the intersection of domain knowledge, forecasting, and visual analytics. As noted, our findings do not fully support the results of previous studies [8, 9, 19, 43]. We find that our participants' forecasts were not significantly different (in terms of accuracy) when compared to the algorithmic forecasts. We also found that participants receiving additional information on the movie business did not perform any better than participants without such knowledge. However, previous studies did not engage the participants with a visual analytics interface. As such, the fact that particpants did not under-perform can be seen as an indicator that a visual analytics approach to forecasting may yield positive results. Although our hypotheses were not supported, our research study remains valuable as it is the first (to our knowledge) controlled study for evaluating human participants during a predictive visual analytics task.

Acknowledgments

This work was supported by the U.S. Department of Homeland Security under Grant Award 2017-ST-061-QA0001 and 17STQAC00001-03-03. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

[1] S. Afzal, R. Maciejewski, and D. S. Ebert. Visual analytics decision support environment for epidemic modeling and response evaluation. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 191–200, Oct 2011.

[2] H. R. Arkes, R. M. Dawes, and C. Christensen. Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37(1):93–110, Feb 1986.

[3] V. Buchanan, Y. Lu, N. McNeese, M. Steptoe, R. Maciejewski, and N. Cooke. The role of teamwork in the analysis of big data: A study of visual analytics and box office prediction. *Big Data*, 5(1):53–66, 2017.

[4] J. Buchmüller, H. Janetzko, G. Andrienko, N. Andrienko, G. Fuchs, and D. A. Keim. Visual analytics for exploring local impact of air traffic. *Computer Graphics Forum*, 34(3):181–190, Jul 2015.

[5] D. Cashman, G. Patterson, A. Mosca, N. Watts, S. Robinson, and R. Chang. Rnnbow: Visualizing learning via backpropagation gradients in rnns. *IEEE Computer Graphics and Applications*, 38(6):39–50, Nov 2018.

[6] M. Cavallo and Ç. Demiralp. Clustrophile 2: Guided visual clustering analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):267–276, Jan 2019.

[7] M. B. Cook and H. S. Smallman. Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors*, 50(5):745–754, 2008.

[8] R. M. Dawes. The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571–582, Jul 1979.

[9] R. M. Dawes, D. Faust, and P. E. Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, Mar 1989.

[10] B. J. Dietvorst. People reject (superior) algorithms because they compare them to counter-normative reference points. *SSRN Electronic Journal*, Dec 2016.

[11] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, Feb 2015.

[12] B. J. Dietvorst, J. P. Simmons, and C. Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.

[13] M. El-Assady, W. Jentner, M. Stein, F. Fischer, T. Schreck, and D. Keim. Predictive visual analytics: Approaches for movie ratings and discussion of open research challenges. In *Proceeding of the IEEE VIS Workshop on Visualization for Predictive Analytics*, Nov 2014.

[14] C. Eroglu and K. L. Croxton. Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, 26(1):116–133, Jan 2010.

[15] P. Federico, M. Wagner, A. Rind, A. Amor-Amoros, S. Miksch, and W. Aigner. The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 92–103, Oct 2017.

[16] R. Fildes and P. Goodwin. Against your better judgment? how organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6):570–576, Dec 2007.

[17] A. Furnham and H. C. Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, Feb 2011.

[18] M. Gleicher. A framework for considering comprehensibility in modeling. *Big Data*, 4(2):75–88, Jun 2016.

[19] W. M. Grove, D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment*, 12(1):19–30, 2000.

[20] F. Gul. A theory of disappointment aversion. *Econometrica: Journal of the Econometric Society*, 59(3):667–686, May 1991.

[21] S. Highhouse. Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, 1(3):333–342, Sep 2008.

[22] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, Sep 2015.

[23] M. S. Hossain, P. K. R. Ojili, C. Grimm, R. Müller, L. T. Watson, and N. Ramakrishnan. Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2829–2838, Dec 2012.

[24] M. Kahng, N. Thorat, D. H. Chau, F. B. Viégas, and M. Wattenberg. GAN Lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):310–320, Jan 2019.

[25] J. Klayman, J. B. Soll, C. González-Vallejo, and S. Barlas. Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79(3):216–247, Sep 1999.

[26] J. Krause, A. Perer, and E. Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1614–1623, 2014.

[27] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 5686–5697, May 2016.

[28] R. Krüger, H. Bosch, D. Thom, E. Püttmann, Q. Han, S. Koch, F. Heimerl, and T. Ertl. Prolix – visual prediction analysis for box office success. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, 2013.

[29] M. Lawrence, P. Goodwin, M. O'Connor, and D. Önkal. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3):493–518, Jan 2006.

[30] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164, Jun 2012.

[31] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):91–100, Jan 2017.

[32] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski. The state-of-the-art in predictive visual analytics. *Computer Graphics Forum*, 36(3):539–562, Jun 2017.

[33] Y. Lu, R. Krüger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski. Integrating predictive analytics and social media. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 193–202. IEEE, 2014.

[34] Y. Lu, F. Wang, and R. Maciejewski. VAST 2013 mini-challenge 1: Box office VAST – team VADER. In *Proceeding of the IEEE Conference on Visual Analytics Science and Technology*, 2013.

[35] Y. Lu, F. Wang, and R. Maciejewski. Business intelligence from social media: A study from the VAST box office challenge. *IEEE Computer Graphics and Applications*, 34(5):58–69, Sep 2014.

[36] Y. Ma, T. Xie, J. Li, and R. Maciejewski. Explaining vulnerabilities to adversarial machine learning through visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1075–1085, Jan 2020.

[37] B. P. Mathews and A. Diamantopoulos. Managerial intervention in forecasting. an empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, 3(1):3–10, 1986.

[38] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, Jun 1998.

[39] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, May 2000.

[40] J. Pearl. Comments on Neuberg's review of causality. *Econometric Theory*, 19(04), Jun 2003.

[41] F. Petropoulos, R. Fildes, and P. Goodwin. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*, 249(3):842–852, 2016.

[42] P. E. Rauber, S. G. Fadel, A. X. Falcao, and A. C. Telea. Visualizing the hidden activity of artificial neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):101–110, Jan 2017.

[43] N. Silver. *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't.* Penguin Press, New York, NY, USA, 2012.

[44] T. Spinner, U. Schlegel, H. Schafer, and M. El-Assady. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074, Jan 2020.

[45] Twitter. Twitter streaming api. `https://dev.twitter.com/`

streaming/overview, 2015. Accessed: 2015-11-24.

[46] S. Van Den Elzen and J. J. van Wijk. BaobabView: Interactive construction and analysis of decision trees. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 151–160, Oct 2011.

[47] S. I. Vrieze and W. M. Grove. Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice*, 40(5):525–531, Oct 2009.

[48] X. Wang, D. H. Jeong, W. Dou, S.-W. Lee, W. Ribarsky, and R. Chang. Defining and applying knowledge conversion processes to a visual analytics system. *Computers & Graphics*, 33(5):616 – 623, Oct 2009.

[49] K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mané, D. Fritz, D. Krishnan, F. B. Viégas, and M. Wattenberg. Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):1–12, Jan 2018.

[50] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, and F. Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 307–317, Mar 2017.