Dear Reviewer 2:

I will respond to each point by quoting each of your questions. Regarding your commentary:

> There is no information regarding the product categories, frequency of possible purchases (daily? weekly? monthly? etc.), and the customer histories in relation with those specific products. For instance, what are the types of products, how frequently is it purchased (with respect to the 3 months period analyzed)? What is the expected impact of TV advertisements for those product categories? For how long have the TV advertisements been running for those products? If a customer has already been buying the same product, what is the expected impact of TV advertisements for that product category, according to findings in the literature, if they exist?

The detailed information regarding the product categories was not included since a wide array of different products are included in the original dataset and it detracts from the main objective of the research and discussion. However, we included a summary of the matched products in the revision in Appendix B.1 Surveyed Products and Figure B.1 of the revised section of this manuscript.

The frequency of purchase is out of the scope of the survey data that we received, so it is not included. The question was "Have you purchased the product in the last month?" for all products. The expected impact of the TV advertisements would certainly be useful to make a measured comparison, but this is private information for each individual company and it is out of the limitations of this study. The lenght of the advertisements being run is not available outside of the span of those 3 months, as per the data we have received. This is also a limitation of this study. We have added these points in section 8. Limitations.

> While "Based on the input vector that is used" is mentioned in the manuscript, no explanation is given regarding the input vector.

The previous input vector was described in section 4.2 Advert Viewing Time Calculation, further shown in Table 3 of the first version of the manuscript. However, I realize it is a bit unreadable, so it is described below:

There were two models used:

- Total Seconds Model: A single feature, the total number of seconds observed for the related advert.
- Day/Primetime Model: Each day is split into two features, Primetime and Non-Primetime. Then the same is added with Weekdays, Weekends, Holidays and Rest days. Total of 22 features.
  - Monday Primetime
  - Monday Non-Primetime
  - Tuesday Primetime
  - Tuesday Non-Primetime
  - Wednesday Primetime
  - Wednesday Non-Primetime
  - Thursday Primetime
  - Thursday Non-Primetime
  - Friday Primetime
  - Friday Non-Primetime
  - Saturday Primetime
  - Saturday Non-Primetime
  - Sunday Primetime
  - Sunday Non-Primetime
  - Weekday Primetime
  - Weekday Non-Primetime
  - Weekend Primetime

– Weekend Non-Primetime

– Holidays Primetime

– Holidays Non-Primetime

– Rest Day Primetime

– Rest Day Non-Primetime

That is to say, in order to answer a request from a different reviewer to add Multi-Layer Perceptron and Logistic Regression, as well as comments received on this review, aditional experiments were done. The design of the feature vector will be different from this in the submitted version, so this is merely explanatory of the first version of the manuscript. In regards to the revised version, the subsections of section 4. Methodology of this manuscript were revised to clarify the experiments further. The input vector design is sumarized as well in 4.1. Experiment Design Overview. We also included a short summary further below.

The changes in the experiments and explanations are reflected in the manuscript. This resulted in a new section structure. An overview of the experiment design was laid out in the new section 4.1 Experiment Design Overview.

> Positive and negative class definitions are given for Product Based models, but not for User Based Models.

Both class definitions apply the same for User Based Models. The Prediction Target is the same, and the positive and negative class are for the Prediction Targets, not for the input vectors.

> Table 5 mentions 200 products, but according to explanations in other places, only the data related to 38 products are used in the experiments. For each user, does it mean that there are only 38 data points? Is the prediction made by using 38 data points for each user? Did each user have a purchase intent for all 38 products? If not, were there less number of data points for some users?

The data from the advert viewing time and the survey aren't particularly linked in the original dataset that we received from the Nomura Research Institute. When relating the 200 product survey results to the adverts run in the span of time that the survey points to, only 36 products ran ads that we had access to the viewing time for the same group of users. In the previous version of this manuscript, 2 more products were used that were not filtered correctly at the time. This was corrected in the new experiments. To answer the question, each user has 36 data points relating to each product, and the prediction is made with this data. The purchase intent is stated (yes or no) for all of these 36 products. All users have the same amount of data points.

> What about the results for other categories? This might provide additional insights and improve the quality. Other categories are listed, but the corresponding results are not given.

The results from other categories were originally thought uninteresting to the purpose of the paper and would clutter the main topic of the study. However, because of the number of experiments after the new design, we included averages and t-test results for all categories.

> What are the sizes for all categories? Are they similar?

The sizes for all categories change between product and product, so the Appendix B.3. Prediction Target Categories was added and a comprehensive detail is shown in Table B.20.

> What are the advertisement frequency and exposure times for different products? Are they similar?
>
> What is the viewing time on a product and user basis (e.g. histogram or other aggregate information, if details cannot be shared)?

The detailed advertisement frequency and exposure times for different products, as well as specific viewing times were not disclosed because of legal limitations with the source of the data, Nomura Research Institute. However, the

summaries of these data were added to the manuscript in Appendix B.2 Advert Exposure and Broadcasting Data, summarized in Figures B.2 and B.3.

> Only a few specific product names are given, different product types are not described and discussed. For an international audience, explanations about the product categories would provide more information.

A general distribution of the matched 36 products in simple categories was added in Appendix B.1 Surveyed Products, shown in Figure B.1. Because of the new design of the experiment, however, specific Japanese product names were removed from the manuscript.

> No explanation is given regarding the distribution of commercial broadcast times. This weakens the primetime related conclusions.

A graph describing the general distribution of commercial broadcast times was added for the 36 products in Appendix B.2 Advert Exposure and Broadcasting Data, in Figure B.2.

> Any descriptions of the setting where data were collected are missing. Only, "A real life environment" is mentioned as a description.

The data given to us by the Nomura Research Institute contains actual television viewing (viewing counted as the television being on and tuned to a particular channel) time from households across Japan. This was clarified further in section 4.4.1. Advert Viewing Time

> "We created different datasets so that the viewing time of one element never repeated in a later one." Is this correct? For instance, weekday includes other daily viewing times such as Monday, Tuesday,... in Table 3.

This text was referring to the fact that total viewing time was included in a separate model altogether. However, this comment is correct and there were repeating data in our first experiments. In accordance to corrections made for other points of this review and from another reviewer, as well as this particular point, the experiment was redesigned and calculations were done again. The new design of the experiment will be stated further down below.

> What does "CM view" refer to? The abbreviation appears in many places. Is it "commercial view"?

This is an expression common in Asian countries, abreviating Commercial to "CM". It was incorrectly used in this paper so we have exchanged all instances to "Advert Viewing".

> The use of the term "Not considered?" in Table 1 is confusing. Instead, it might be better to state it as "Yes/No" to be consistent with other entries in the table. "Not considered" sounds as if the customer did "not consider" the product for purchasing.

We have corrected the tables in this revision accordingly. Because of new tables introduced in the revision of the manuscript, these tables are now Tables 3 and 4.

> In the manuscript: "observing the relation between the previously explained data and the effctiveness of commercial adverts on television" -¿ It is not clear which data are referred to by the term "previously explained data".

"Previously explained data" was referring to the Purchase Intention and Actual Purchase and their different categorizations, as they are the direct previous section of the study. However, to avoid any confusion, we have revised this statement. Because of the new experiment design, the section is now 4.4.1. Advert Viewing Time, whereas before it was 4.2. Advert Viewing Time Calculation in the previous manuscript.

> Regarding the choice of the machine learning methods used, not much explanation/justification is provided.
>
> Linear kernel is used for SVC in the study, but did not prove to be successful in predictions. Such issues were left unaddressed: Was another approach considered for the prediction model? Would the use of nonlinear kernel be useful? Would it be possible to use another method? Would logistic regression be applicable? etc.

The reason we used SVM and XGBoost is that they are considered well performing machines in the machine learning field for the size of the available data that we had. Other methodologies are generally thought perform better under the assumption that databases on a different order of magnitude are used. However, this was also a point of concern for another reviewer, so we have added another analysis method, logistic regression, in order to have a better scope for comparing results. We have also added experiments using demographic data for comparison and control.

Our new experiments will be done with 3 methodologies:

- SVM
- XGBoost
- Logistic Regression

Regarding the non-linear kernel. The experiment was actually attempted, but the non-linear versions of SVM algorithms are easy to train into infinite loops when the predictions can't be improved. Our attempt resulted in this manner, so it was not included in the paper.

> With respect to the choice of alternative methods, no explanations are provided other than mentioning: "as well as using different machine learning algorithms, which weren't considered because of requiring bigger datasets." in the last sentence of the manuscript.
>
> It is not clear what is meant by "requiring bigger datasets".

This refers mostly to the newer and state of the art Deep Neural Networks. Particularly for Deep Learning, datasets with orders of magnitude much larger (for example, 3,000,000 users instead of 3000, and hundreds or thousands more products instead of 36) are generally thought to be necessary to perform better. This explanation was added to the manuscript in section 8. Limitations.

> Comparisons are made against other experiments carried out in controlled environments in the literature, and the findings of those studies are criticized due to the argument that they do not reflect real life settings. Because there are many factors affecting the customer purchase behavior, it may be difficult to isolate the influence of TV advertisements by itself, in an uncontrolled setting. This may be related to the poor prediction results, aside from possibly not using a more suitable model.

In order to correctly make comparisons with previous experiments in controlled environments, as well as to include a number of factors that could increase the chances of improving the model prediction resutls, even in a real life setting, we have decided to redesign the experiment (and input vectors) as follows.

- Product Based Models:
  - Advert Viewing Time
  - Advert Viewing Time, Demographics, Purchase Intention
  - Demographics, Purchase intention
- User Based Models:
  - Advert Viewing Time
  - Advert Viewing Time, Demographics, Purchase Intention
  - Demographics, Purchase intention

* Of course, removing the Purchase Intention from the feature vector when it is the Prediction Target

With 3 methodologies:

- SVM
- XGBoost
- Logistic Regression

The models and features defined as follows:

- Advert Viewing Time:
  - Weekday Configuration:
    - Monday
    - Tuesday
    - Wednesday
    - Thursday
    - Friday
    - Saturday
    - Sunday
  - Weekday and Time Slot Configuration:
    - Monday Primetime
    - Monday Non-Primetime
    - Tuesday Primetime
    - Tuesday Non-Primetime
    - Wednesday Primetime
    - Wednesday Non-Primetime
    - Thursday Primetime
    - Thursday Non-Primetime
    - Friday Primetime
    - Friday Non-Primetime
    - Saturday Primetime
    - Saturday Non-Primetime

        * Sunday Primetime
        * Sunday Non-Primetime

- Demographics:
    - Age:
        * 18 to 25 years old
        * 26 to 35 years old
        * 36 to 45 years old
        * 46 to 55 years old
        * 56 or older
    - Sex:
        * Male
        * Female
    - Marital Status:
        * Single
        * Married
        * Divorced or Widowed
    - Parental status:
        * Parent
        * Not a Parent
    - Income Bracket:
        * Not disclosed
        * No Income
        * Under 1,000,000 yen
        * From 1,000,000 yen to 2,000,000 yen
        * From 2,000,000 yen to 3,000,000 yen
        * From 3,000,000 yen to 4,000,000 yen
        * From 4,000,000 yen to 5,000,000 yen

        * From 5,000,000 yen to 6,000,000 yen
        * From 6,000,000 yen to 7,000,000 yen
        * From 7,000,000 yen to 10,000,000 yen
        * From 10,000,000 yen to 15,000,000 yen
        * From 15,000,000 yen to 20,000,000 yen
        * Over 20,000,000 yen

- Purchase Intention:
    - January Survey response
    - March Survey response

This corrects for several points. Without repeating any day of advert viewing time, we can safely assume that the input vector is not dependant upon itself in a mathematical manner. It also introduces data points that are common in previous literature, such as age, sex, marital status and income brackets, as well as proving as a comparison point between the results of different models including or excluding the advert viewing time. If models without the advert viewing time were to perform better than those that include it, constantly and with statistically significant differences across experiments, this could mean that the external factors have a determinant influence on purchase behavior. If it were the other way around, it could prove that adverts are having an effect on purchase behavior. In the revision of this manuscript, we performed these experiments and then compared their average performance with t-tests, to see which variables produce a change in performance. We have detailed this across several sections of the revised manuscript, including the abstract, 2. Research Objective, 4. Methodology and 7. Discussion. In summary, we found that advert viewing time based models performed consistently lower in predicability in comparison with models that include demographic data. In addition to this, we found that the difference in predictability between demographic data models and models that include both demographic data and advert viewing time data was not significant in all cases for actual purchase behavior.

It is reported in the manuscript: "Based on our low results for any prediction of Actual Purchase, we concluded that there must be other factors that are more strongly tied to the customer's purchasing behavior." Due to numerous factors affecting the purchase decision, isolating the impact of commercial view time is a difficult problem. This may be considered and handled by experiments in controlled settings, as reported by some studies in the literature. While the large number of customers and related data are valuable, lack of access to a controlled environment makes it difficult to assess properly the impact of TV advertisements alone.

With the new design of the experiments explained above, in this revision we attempt to isolate the impact of commercial view time in comparison to other models that don't include it, but include data largely used in previous literature.

"Points left to research in future work are a deeper analysis of the predictable customers, looking for similarities or clusters within this class, as well as using different machine learning algorithms, which weren't considered because of requiring bigger datasets." -¿ Such findings would improve the level of contribution significantly. However, currently it is rather limited, and not up to its potential in terms of possible insights hidden in the data. (The data collected may have significant potential.)

Although his study certainly has its limitations in size of data and freedom of use, we redesigned the experiments and performed a series of tests that could indicate doubt for the current television advertising strategies, and revised the manuscript accordingly. Our new results consistently indicated that advert viewing time was having no effect in purchase behavior as well. We believe that this doubt alone is a contribution for future research to be continued in this field that is currently stagnated.

In its current form, the manuscript is not at an acceptable qual-

> ity. Lack of some explanations and discussions regarding the used approach makes it rather difficult to assess the level of contribution. Accordingly, the opinion of the reviewer is inclined toward rejection of the paper, unless a major revision can be carried out in order to address the issues given above.

We have revised the manuscript and redesigned the experiment in accordance to this and other reviewers, obtained new solid results that are consistent with our original views, and hope that it is appropriate in its revised form for your consideration. A number of new sections in response to this review, as well as a revision to the related work, results, discussion and conclusion sections of this manuscript were necessary after these experiments. We ask that you consider the major revision addressing the previous issues.