Exploring the Expedia Recommendation Algorithm

Group 35: Max Ming Yi Koh (1007972785) and Kevin Le (1007952805)

March 31, 2022

General Motivations and Dataset Description

- Expedia generates revenue by reselling bookings (purchased in bulks at discounted pricing) or charging commissions from hoteliers.[1]
- Thus, an effective recommendation algorithm that recognizes consumer needs is crucial to improve user experience and optimize the profitability of the business.
- We formulate 3 research questions that answer these broad scoped questions:
  1. How effective is Expedia's recommendation algorithm?
  2. What are some factors that affect purchasing decisions of consumers?
- The investigation uses a dataset of 1,000 Expedia user searches along with certain variables related to users or the top 3 recommended properties between June 1^{st}, 2021 and July 31^{st}, 2021.
- Unless restated, you may assume the sample of the 3 research questions consists of the users that made these 1,000 Expedia searches while the population consists of all Expedia users who made a search between June 1^{st}, 2021 and July 31^{st}, 2021.

Data Summary

Below is a table of variables used in the 3 research problems. Note that {n} is a placeholder for integers 1, 2, or 3.

Variable	Description
is_trans{n}	whether the consumer transacted the n^{th} displayed property within 180 minutes of a user search
is_drr{n}	whether the n^{th} displayed property is discounted
num_clicks{n}	number of clicks for the n^{th} displayed property within 180 minutes of a user search
checkin_date	stay start date
checkout_date	stay end date
adult_count	number of adults on the trip
child_count	number of children on the trip

Research Question 1 - Introduction

Research Question: What is the proportion of consumers between June 1^{st}, 2021 and July 31^{st}, 2021 that purchase one of the top 3 recommended properties within 180 minutes of a search?

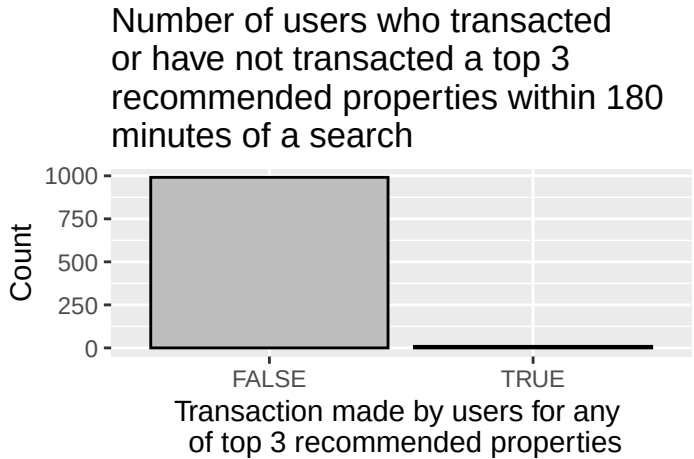
Research Motivation:

- The proportion of users who purchase a top 3 recommended property is a good metric to measure the effectiveness of Expedia's recommendation algorithm.
- For instance, a higher proportion may imply the algorithm is effective at recognizing user needs while a lower proportion may imply the algorithm is less effective at recognizing user needs.

Research Question 1 - Data Wrangling

- We applied the `select()` function to obtain the required variables, namely is_trans1, is_trans2 and is_trans3.
- Using is_trans1, is_trans2 and is_trans3 columns, we applied the `mutate()` function to create a new variable named trans_made which indicates whether any transactions were made by a user within 180 minutes of his or her search.
- trans_made is set to TRUE if a transaction is made within 180 minutes of a user search. Otherwise, trans_made is set to FALSE.

Research Question 1 - Data Visualization



This figure shows that 9 out of 1,000 consumers in the sample purchased a top 3 listing recommended by Expedia within 180 minutes of their search.

Research Question 1 - Statistical Analysis

- We assume that the sample is representative of the population in order to perform bootstrapping.
- By resampling from the sample of 1,000 users for 3,000 repetitions and choosing a 95% confidence level, we find the confidence interval for the proportion of users between June 1^{st}, 2021 and July 31^{st}, 2021 who purchased a top 3 recommended property within 180 minutes is (0.004, 0.015).
- A confidence level 95% means that 95% of confidence intervals generated in a similar manner (i.e. resampling from the sample of 1,000 user data 3,000 times) will capture the true proportion of consumers between June 1^{st}, 2021 and July 31^{st}, 2021 who purchased a top 3 recommended property within 180 minutes.
- The width of the confidence interval (i.e. $0.015 - 0.003 = 0.012$) is very narrow. So, we expect the true proportion to be very similar to the estimate made.

Research Question 2 - Introduction

Research Question: Do the mean number of clicks received within 180 minutes of a user search differ for top 3 recommended discounted and non-discounted listings between June 1^{st}, 2021 and July 31^{st}, 2021?

Hypothesis: The population contains top 3 recommended properties for each search between June 1^{st}, 2021 and July 31^{st}, 2021.

$H_0 : \mu_{no\ discount} - \mu_{discount} = 0$

$H_1 : \mu_{no\ discount} - \mu_{discount} \neq 0$

where $\mu_{no\ discount}$ and $\mu_{discount}$ are the mean number of clicks for non-discounted and discounted top 3 listings respectively between June 1^{st}, 2021 and July 31^{st}, 2021.

Research Motivation:

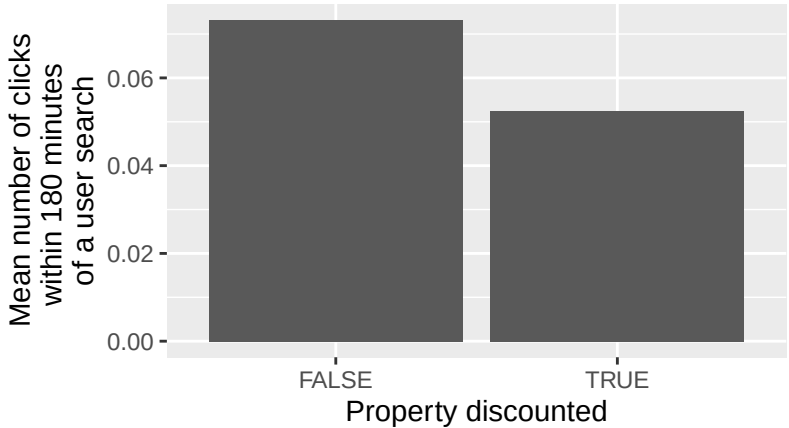
- Hypothetically, if users are more interested in discounted listings, Expedia may increase the number of discounted listings to attract more consumers.

Research Question 2 - Data Wrangling

- We applied the `select()` function to obtain the required variables, namely num_clicks1, num_clicks2, num_clicks3, is_drr1, is_drr2 and is_drr3.
- To tidy the data (due to change in population), we ran a for loop to reshape the tibble to have rows which represent listings instead of user search results. The new tibble has columns num_clicks and is_drr.
- num_clicks represents the number of clicks received by the property within 180 minutes of a user search.
- is_drr represents whether the listed property is discounted.
- We applied the `group_by()` and `summarise()` functions to obtain the mean number of clicks received within 180 minutes of a search for discounted and non-discounted top 3 listings.

Research Question 2 - Visualization

Average number of clicks within 180 minutes of a user search for discounted and non−discounted top 3 properties



Research Question 2 - Statistical Analysis

- The calculated test statistic from the dataset for the difference between the mean number of clicks received within 180 minutes of a search for non-discounted and discounted top properties is -0.0208.
- This means that within the sample, discounted top 3 properties get 0.0208 less clicks than non-discounted properties on average within 180 minutes of a user search.
- After running 5,000 simulations under the assumption that the null hypothesis is true (by shuffling the label id_drr indicating whether a property is discounted), the p-value is found to be 0.0498.
- Since the p-value is between 0.01 and 0.05, there is moderate evidence against the null hypothesis which states that the mean number of clicks for discounted and non-discounted properties received within 180 minutes of a user search is the same.

Research Question 3 - Introduction

Research Question: Do the number of adults and children on a trip affect the length of travel?

Hypothesis:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

where β_i is a slope coefficient for the linear regression model where the predictors are the number of children and adults and the response is the length of travel.

Research Motivation:

- If a correlation exists between the type and number of travelers and length of travel, Expedia may want to consider this correlation when booking properties in bulk.
- For instance, larger properties which can fit more people may be booked for longer interval of time if the correlation is found to be positive.

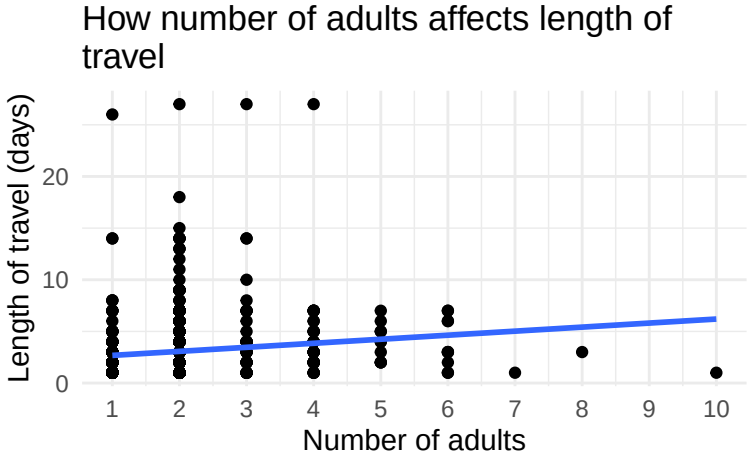
Research Question 3 - Data Wrangling

- We applied the `select()` function to obtain the required variables, namely adult_count, child_count, checkin_date, checkout_date.
- We applied the `mutate()` function to create a new variable called travel_length.
- travel_length represents the time difference between checkout_date and checkin_date in days.
- To split dataset into training and testing when performing linear regression, we applied the `rowid_to_column()` function to add a unique identifier to each row in the tibble.

Research Question 3 - Statistical Analysis

- We created several linear regression models which differ by number of predictors (e.g. 1 predictor versus 2 predictors) and whether predictor variables interact.
- By picking the "best" model based on prediction accuracy, low signs of overfitting and simplicity, we find that the "best" model is a simple linear regression model (a model with 1 predictor) that uses the number of adults to predict the length of a trip.
- The extremely small p-value of 5.55×10^{-4} suggests there is very strong evidence against the null hypothesis that there is no relationship between the number of adults on a trip and the length of a trip. Thus, we reject the null hypothesis.
- The calculated slope of 0.4258 implies that for each additional adult, the length of trip increases by 0.4258 days on average.
- However, the R^2 value of 0.0148 for the linear regression model implies only 1.48% of variability in the length of trip is accounted by the number of adults on the trip.

Research Question 3 - Visualization



This is a figure of the "best" linear regression model which shows the association between the number of adults on a trip and the length of travel.

Limitations

- We assume that data filled in by users is accurate. Inaccurate data entries (e.g. typos) may skew the statistical results.
- We assume that the sample is representative of the population. Otherwise, statistical models (especially bootstrapping) will not yield accurate results.
- For research questions 1 and 2, is_trans{n} and num_clicks{n} only measure user events like transaction or clicks for 180 minutes after a search. This time interval limitation carries over to the statistical analysis.
- For research question 2, transactions are a better metric of consumer interest compared to clicks received by a listing. However, only 9 properties are transacted out of the 3,000 properties. So, we used the number of clicks as the metric.
- For research question 3, we intended to research how the number of adults, children and infants affect the length of travel. However, only 9 data points had infants in their travel group. So, we removed the number of infants as a potential predictor.

Overall Conclusion - Looking Ahead Part 1

Here are some closing thoughts for each research question.

For the **research question 1**,

- The 95% confidence interval for the proportion of top 3 recommended listings transacted is (0.004, 0.015).
- The bounds of the interval are low which implies the recommendation algorithm has room for improvement in terms of increasing the proportion of users who make a transaction (within 180 minutes).
- We recommend Expedia to invest in research and development of recommendation algorithms as such algorithms have potential of directly improving the profitability of Expedia.

Overall Conclusion - Looking Ahead Part 2

For the **research question 2**,

- It was found that there is weak evidence against the null hypothesis that the mean number of clicks for discounted and non-discounted properties is the same.
- We recommend Expedia to perform A/B testing with 2 variations of the website which differ by the number of discounted listed properties.[2]
- This allows Expedia to find out how user consumption behaviours change based on change in the number of discounted properties.
- Knowing this may help change the recommendation algorithm (in terms of whether it recommends more discounted or non-discounted properties) to suit user preferences.
- At the same time, understanding consumption behaviours may lead to the development new marketing strategies in terms of discounting more or less properties to maximize profit.

Overall Conclusion - Looking Ahead Part 3

For the **research question 3**,

- It is found that as the number of adults increases, the length of travel increases.
- Expedia can consider the linear regression model when purchasing properties in bulk or advising hoteliers regarding the expected intervals of booking.
- For instance, when booking larger properties (which can fit more people) for resell, Expedia should book these properties for longer period to reduce chances of them not being transacted.
- However, the low R^2 value of 0.0148 of the model implies there are other variables (apart from number of adults) that explain the variability in travel length.
- To get a more "complete" model, Expedia may choose to continue to explore how other variables (including those that do not appear in the current dataset) affect length of travel.

Citations

1. https://www.nasdaq.com/articles/how-expedia-makes-most-its-money-2017-08-28
2. https://vwo.com/ab-testing/