Homework 1
Hengxu Lin (hl3541)
October 10, 2022

1 Problem 1

1.1 Task Introduction

Expedia, the largest online travel company in the world, seeks to maximize hotel bookings through its websites. It can take certain actions to influence customer decisions, such as adjusting the number, order and types of hotels displayed. The actions can be based on information available before and after a customer starts a booking session (an episode).

The goal of this problem set project is to utilize the Expedia dataset, implement a probabilistic model (Bayesian Logistic Regression) to predict the click-through actions of the consumers.

1.2 Dataset

- **Description**. The Expedia dataset is a private dataset obtained from Wharton Customer Analytics. The data is comprised of a search transaction dataset and a clickstream dataset. I use the search transaction dataset in this project, which includes user search record per hotel on Expedia.
- **Selected Features.** I manually select 16 categorical and numerical variables which is related to consumer clicks in Appendix Table 3.1, this yields 1,359,595 observations. The features include searching criteria, search results, price, rating, promotion, hotel characteristics and the dependent variable is whether a hotel is clicked.

1.3 Data Analysis

- **Feature Engineering.** I check the feature distribution to decide whether the feature should be discarded, normalized or further feature crafted. I transformed all integer features into a uniform distribution $U(0, 1)$ with standard deviation of $\sqrt{1/12}$ since this features (ranking, star) is close to a uniform distribution and re-scaling could preserve meaningful coefficients. I encode all categorical features to dummy variables with the first class as a reference. Finally I check the distribution of *price, room capacity, review counts, review rating* and find that price follows a log normal distribution and others could be transformed to a standard normal with z-score normalization. The crafted feature (excluding dummies) distributions is displayed in Figure 3.2.

- **How clicks related to the hotels display?** I analyze the hotel position in Expedia website, which is comprised of page and the location within the page (each page is limited by 15 hotels). (1) Among all clicked and unclicked hotels, consumers tend to choose hotels in page 1 and 2 with the location in the front of a page 3.2. This shows an effective recommendation of Expedia website. (2) Consumers click hotels with more reviews, promotion and higher star rating. This could be inferred from the correlation scatter plot from Figure 3.2.

1.4 Bayesian Logistic Regression

- **Log-Likelihood** The objective functions is the log likelihood of Bayesian Logistic.

$$LL = \log p(\beta|\mathbf{x}, \mathbf{y}) = -\frac{\lambda}{2}\sum_{k=1}^{p}\beta_k^2 + \sum_{i=1}^{n}(y_i \log\sigma(\beta x_i) + (1 - y_i)\log\sigma(-\beta x_i))$$

- **In/out-sample split.** I shuffle the data and split it with a 0.8:0.1:0.1 ratio with respect to training/validation/testing.

- **Bayesian prior.** I choose priors range from $\beta = 0.5$ to $= 2$ since the feature are normalized thus the coefficient scale should be normal. Since the feature are standardized, I do not include an intercept hence I need not specify prior for the intercept. Figure 1.4 shows the prior influence on the Log Likelihood (LL) (y-axis) which is maximized as training epochs increase (x-axis). (1) A prior of $\lambda = 2$ achieved highest LL and fairly low penalty value. (2) If λ is high (i.e. the regularization is weak), the prior has little effect on optimizing LL.

- **Hyper-parameter.** Figure 3.3, 3.3 shows that a too large or small batch size would make penalty undesirable and a small but mild learning rate is suitable for the data. I present results below based on a best hyper-parameter tuning experiment on validation dataset.

- **Coefficients.** The coefficients report the change of log odds $\log\frac{p}{1-p}$ corresponding to a percentage change in feature. As we can see in Table 3.4, higher rating, lower price, upper position and promotion correlate to higher click odds. Interesting stuffs are traveller's reviews do not show great impact on consumer choices.

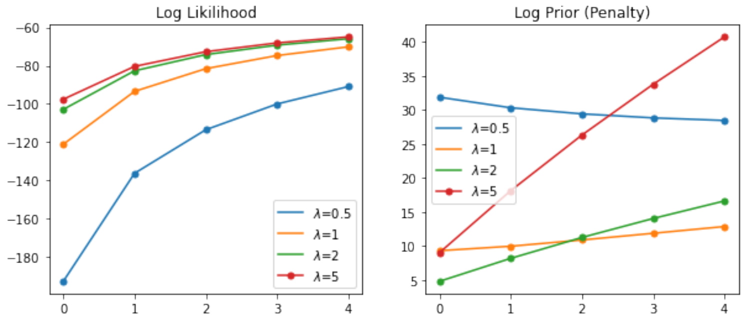

Figure 1: Prior of β Influence

1.5 Classification Reports

Table 1.5 presents the classification results on testing dataset. The clicked sample only accounts for 1.61% of the dataset, but the Bayesian Logistic precision is 2%, which demonstrates that the informative features are captured by the model.

	precision	recall	f1-score	support
unclicked	0.984	0.881	0.930	133729
clicked	0.020	0.145	0.035	2231
accuracy			0.869	135960
macro avg	0.502	0.513	0.483	135960
weighted avg	0.968	0.869	0.915	135960

2 Problem 2

2.1 Variables in the data. (Provided as above)

2.2 Latent variables.

Based on these experiment results I would introduce a latent variable which is related to traveller's reviews. (1) Coefficient results shows these reviews are marginal. There could be a confounding variable explain the relation between the traveller reviews and rating stars. (2) A special prior could be applied to the latent variable so as to control the scale.

2.3 Research question.

- How does the list of hotels displayed to an Expedia user affect the conversion rate for that user? (a) Does the listing order matter? If so, how? (b) Do users respond better to listings that are geographically diverse or diverse in price/quality? (c) Which listing strategies seem to encourage users to continue to refine their criteria versus dropping out? Is there value to looking at the sequence of searches a particular users does?
- How should the search results be optimized to maximize conversion?
- Is there value in tailoring the search results based on other characteristics of the user session? If so, which characteristics are most important? (a) Should results be tailored based on which channel (search engine, tripadvisor.com, etc.) the customer arrived to Expedia through? If so, how? (b) Should results be revised based on which hotels the users looking at in more detail? If so, how?

3 Appendix

3.1 Feature Description

Name	Data Type	Description
TRVL_PRODUCT_ID	int, categorical	Reference for the type of product searched
SRCH_RM(AUDLT)_CNT	int, categorical	Number of rooms (adults) searched for
SRCH_MIN_STAR_RTG	int, categorical	Minimum star rating
TOTL_RESULTS_CNT	int	Number of hotels matching search request
SRT_TYP_ID	int, categorical	The sort type for the search results
RESP_PAGE_POSITION	int	Hotel position within a page
LODG_RESP_PAGE_DESC	int	Hotel in which page
REGIONTYPEID	int	Hotel region
Room_Capacity	int	Count of rooms in the hotel
RESP_STAR_RTG_VALUE	float	Expedia rating of hotel
TRVLR_REVIEW_AVG_RTG	float	Consumer rating of hotel
TRVLR_REVIEW_CNT	int	Count of reviews
IS_PROMO_FLAG	int, categorical	Whether is in promotion
PROPERTY_AVAIL_STATE	int, categorical	Whether is available
LOWST_AVG_PRICE_AMT	float	Hotel price
Viewed_detail (Response Variable)	int, categorical	Whether the hotel is clicked and viewed

3.2 Visualization of feature distribution

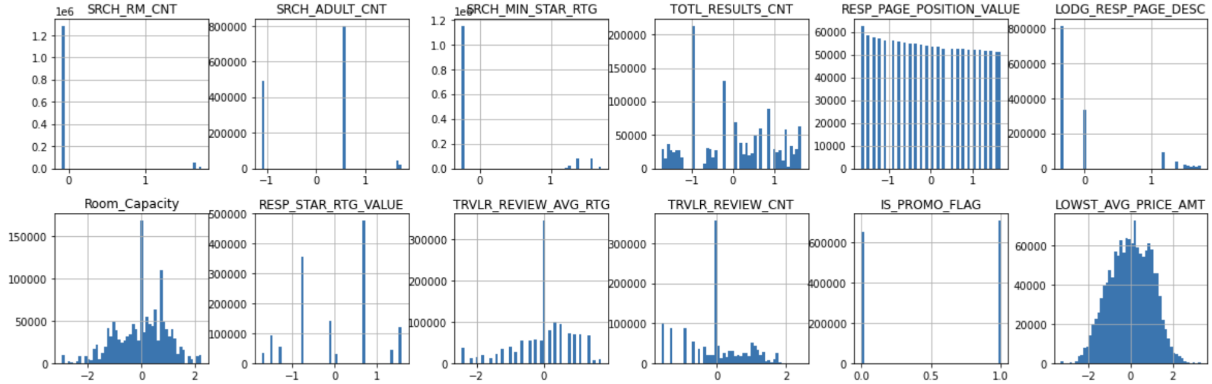

Figure 2: Feature Distribution

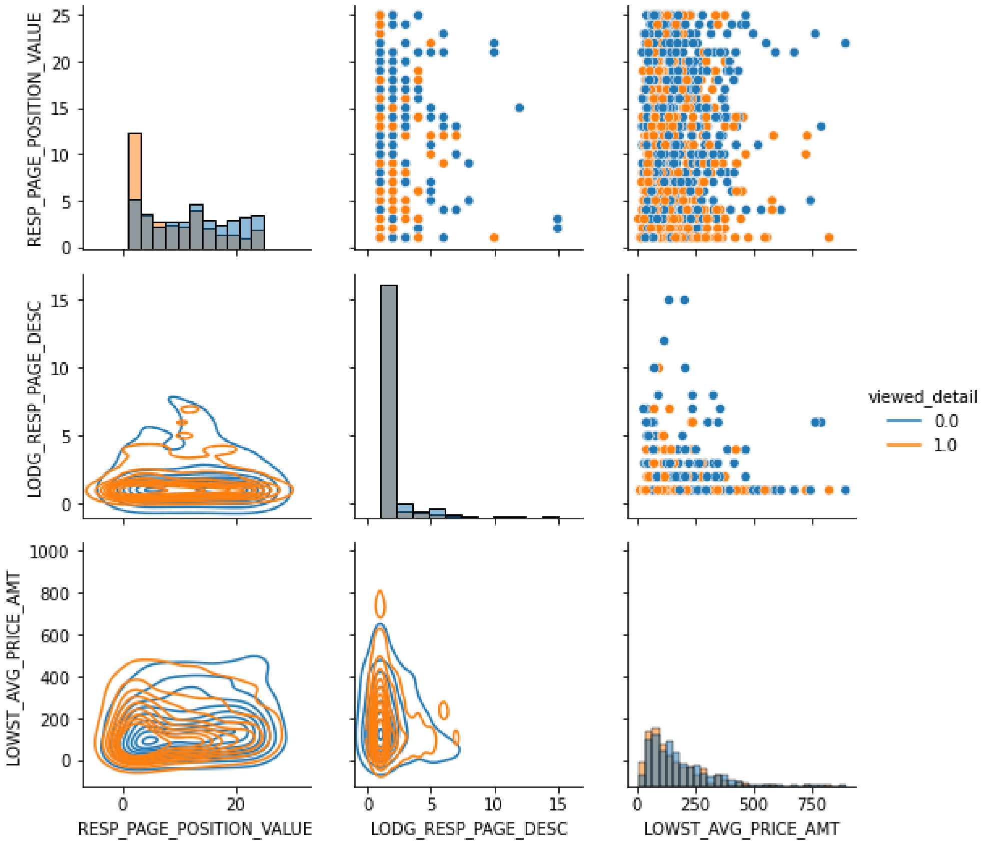

Figure 3: Front position and page are more clicked

3.3 Hyper-parameter Influence

3.4 Coefficients

3.5 Results

	coef
SRCH_RM_CNT	1.167142
SRCH_ADULT_CNT	-0.026331
SRCH_MIN_STAR_RTG	0.279388
TOTL_RESULTS_CNT	0.145840
RESP_PAGE_POSITION_VALUE	-0.328955
LODG_RESP_PAGE_DESC	-0.117146
Room_Capacity	-0.030231
RESP_STAR_RTG_VALUE	0.202469
TRVLR_REVIEW_AVG_RTG	-0.141273
TRVLR_REVIEW_CNT	-0.115919
IS_PROMO_FLAG	-3.947947
LOWST_AVG_PRICE_AMT	-0.548281
TRVL_PRODUCT_ID_3.0	-2.722067
TRVL_PRODUCT_ID_6.0	-2.062712
TRVL_PRODUCT_ID_8.0	-4.806024
TRVL_PRODUCT_ID_9.0	-0.488276
TRVL_PRODUCT_ID_11.0	-2.931265
TRVL_PRODUCT_ID_20.0	-2.355393
TRVL_PRODUCT_ID_25.0	-0.757699
SRT_TYP_ID_1.0	3.155360
SRT_TYP_ID_2.0	0.597196
SRT_TYP_ID_4.0	-1.418002
SRT_TYP_ID_6.0	-1.731138
SRT_TYP_ID_10.0	-1.640827
SRT_TYP_ID_12.0	-3.166724
REGIONTYPEID_14.0	1.406878
REGIONTYPEID_18.0	-1.788658
REGIONTYPEID_19.0	-1.230722
REGIONTYPEID_21.0	-3.111839
REGIONTYPEID_24.0	4.296992

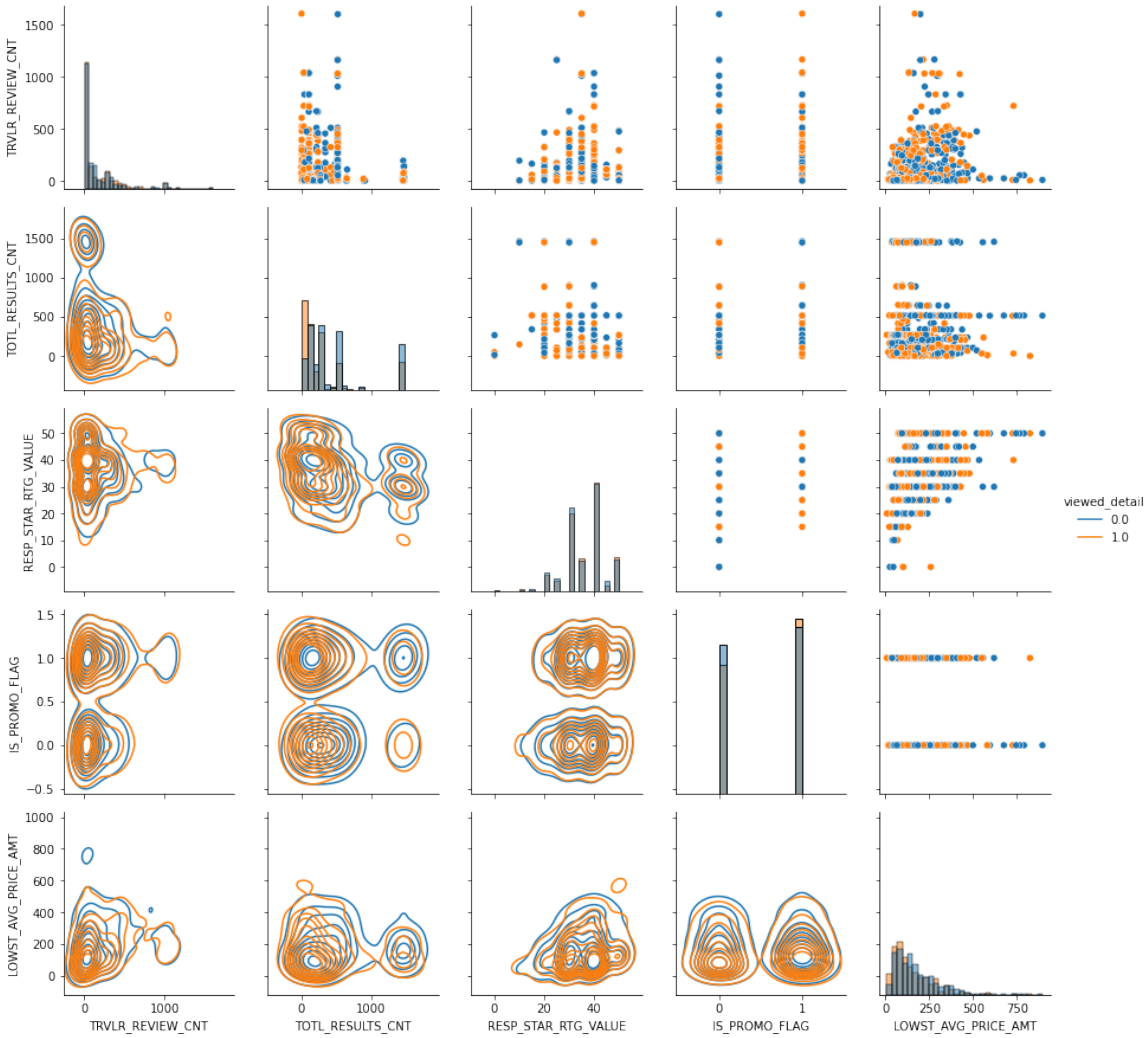

Figure 4: Features are correlated differently in clicked/unclicked search

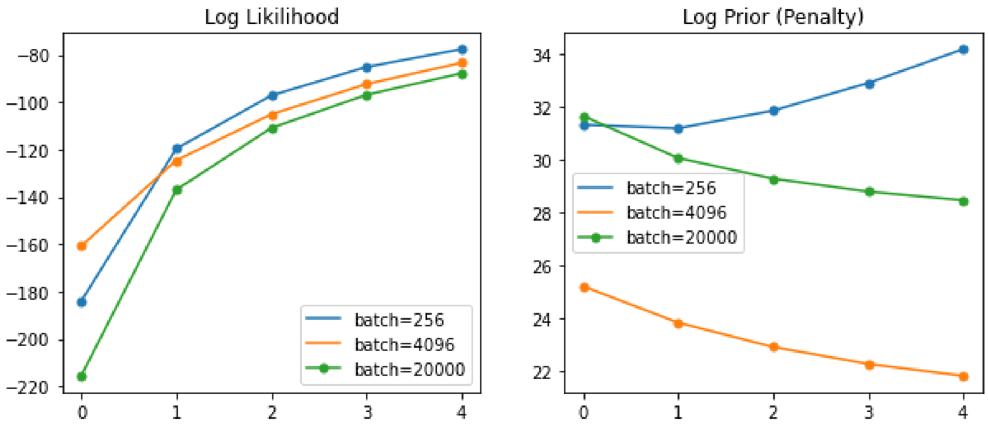

Figure 5: A medium batch size is best

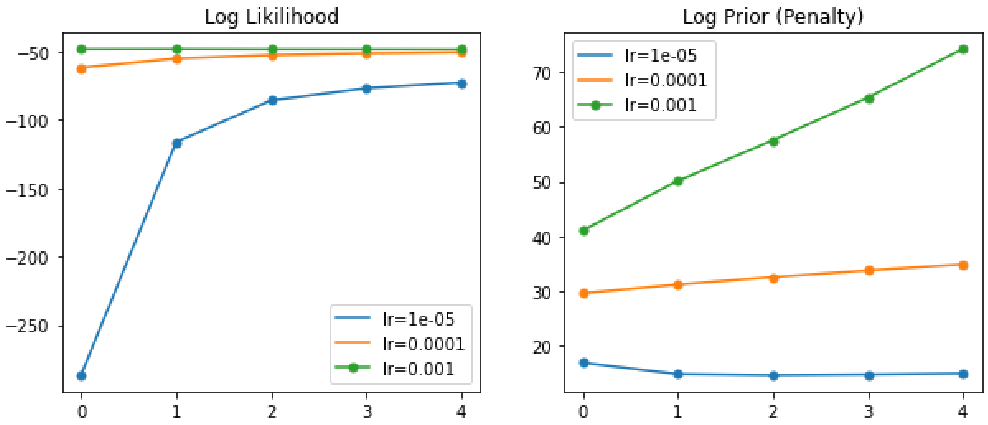

Figure 6: A small learning rate is best

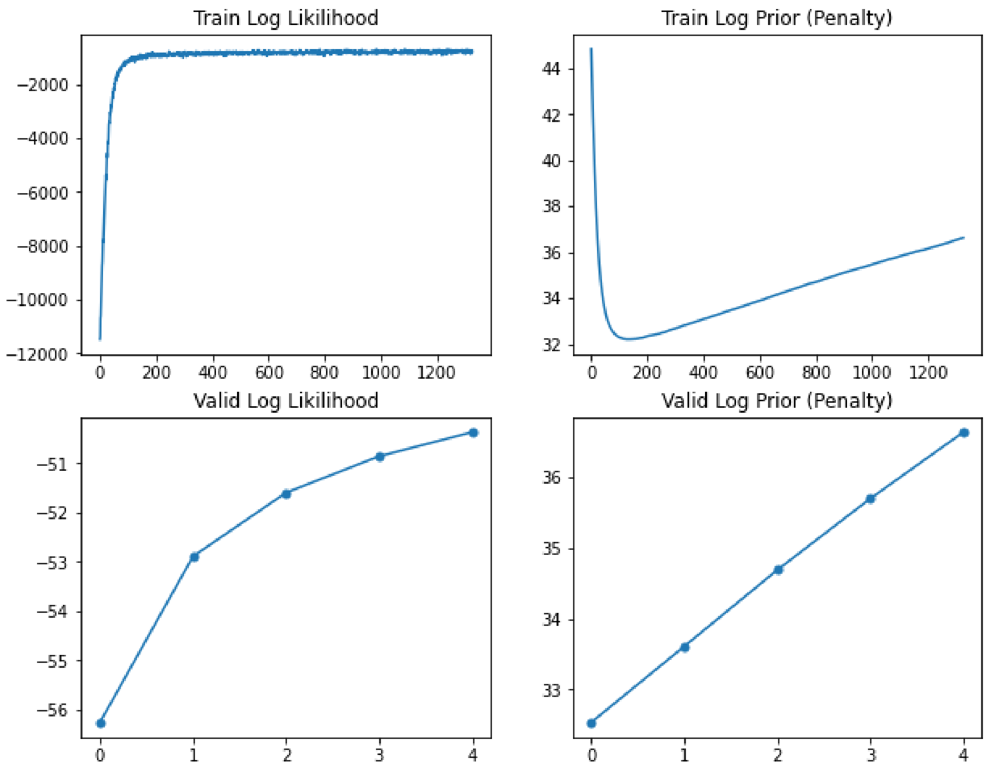

Figure 7: Training and Validation Curves

```
              precision    recall  f1-score   support

           0      0.984     0.881     0.930    133729
           1      0.020     0.145     0.035      2231

    accuracy                          0.869    135960
   macro avg      0.502     0.513     0.483    135960
weighted avg      0.968     0.869     0.915    135960
```

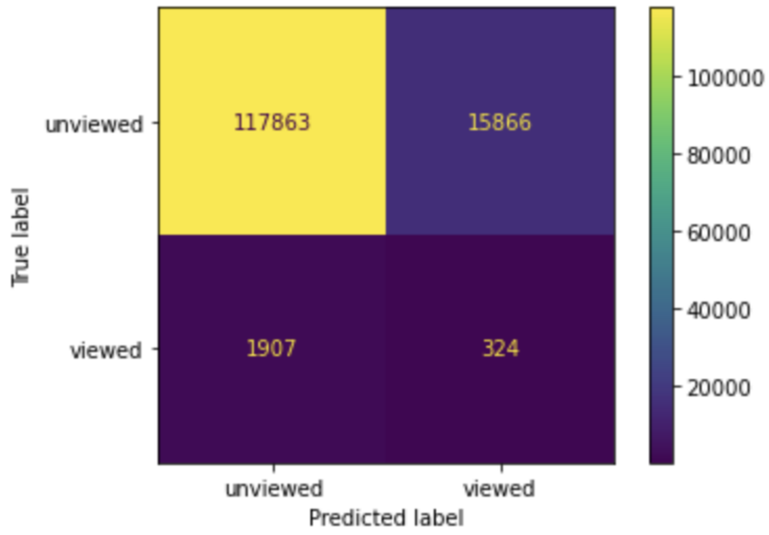

Figure 8: Classification Report