CVR Prediction in Sponsored Search using Logistic Regression

Jiachao Fang, Yuchong Xiao

Introduction

Sponsored search is one of the most popular marketing approaches in todays internet ecosystem. Merchants set several keywords for their products, and these products will be presented to the customers who search these keywords. Conversion rate (CVR) is defined as the probability that the user finally purchases the advertised action. In sponsored search, if items with higher CVR are displayed, it is much easier for customers to find the satisfied product, which improves the user experience during online shopping. On the other hand, more purchasing behavior is preferable to the advertisers as the return from the marketing budget. In summary, estimating the CVR accurately facilitates the search engines to reach more potential customers, increases the return on investment (ROI) for advertisers, and improves the user experience during online shopping.

Given the advantages to both merchants and customers, sponsored search has become a popular research area. Researcher has been working on different approaches to improve sponsor research. Most research focuses on click-through rate (CTR), the probability that the user clicks the advertised item. However, with accessibility of more user data and higher capability of computing, industries have started to make more effort to improve CVR directly, since CVR is directly associated with ROI. Driven by industries, researchers also start to shift their focuses from CTR prediction to CVR prediction. However, the research on CVR prediction is still in an early stage. Only limit research on CVR prediction can be found and most are adapted from former research on CTR.

Due to limit research on this area, we decided to start it in a simple way. Our research on CVR prediction focusesrf on using simple models, such as multivariate regression and logistic regression. Comparing to CTR, CVR involves in a more complicate process, which also depends on web pages after clicking certain links and thus more information should be considered. Our dataset is real transaction data from taobao.com, which is China's largest e-commerce platform, and the data is provided by Alibaba for IJCAI-18 (Alimama Sponsored Search CVR Prediction Contest). Provided dataset consists information of query, user, advertisement, context and shop. We focus our research on feature engineering due to large quantity of information given.

Related Works

Logistic regression In statistics, logistic regression is a regression model of the relationship between a group of independent variables and a binary dependent variable (Stoltzfus 2011). While linear regression is used to generate continuous outcomes, researchers often employ logistic regression to analyze binary outcomes, such as mortality. Logistic regression has been applied in various disciplines, especially in the health-related area. Back to 1989,(Charlton and Blair 1989) built logistic regression models to investigate the factors influencing individual propensity to smoke. (Merlo et al. 2006) used multilevel logistic regression to understand whether residents would consult private physicians based on the health survey data. Researchers also adopt the logistic regression techniques to analyze the effects of advertising. (Ansolabehere et al. 1994) presented logistic models to estimate whether voting intention is statistically dependent on positive and negative advertising. In our research, we choose a multivariate logistic regression model to predict whether a customer will purchase the advertised products, and identify the factors that influence customers decision making.

Feature selection Feature selection has become extremely vital when it comes to data analysis with massive dataset. Guyon and Elisseeff (Guyon and Elisseeff 2003) articulated three objectives of feature selection: improving prediction accuracy, saving time and cost and interpreting data better. For variable subset selection, wrappers, filters and embedded methods are most frequently used. Appropriate feature representations lay the foundation for algorithms to perform efficiently. Most frequently-used feature selection methods include clustering; dimensionality reduction (PCA and LDA); linear transformations (Fourier, Hadamard) (Guyon and Elisseeff 2003). Feature selection is important to prediction of CVR that we have a large dataset, which contains potential features to be extract. In our research, we use filters and PCA to selection feature and conduct dimensionality reduction. Linear transformations will also be considered for further research design.

CTR/CVR prediction Sponsored search advertising is a great way to guide traffic and increase business. To understand the CTR/CVR is central to the multi-billion-dollar online advertising industry. (Richardson, Dominowska, and Ragno 2007) presented a logistic regression model to estimate the CTR for new ads. (Ciaramita, Murdock, and Plachouras 2008) used online multilayer perception learning model for large scale data retrieved from Web search engines. (Li et al. 2015) utilized learning-to-rank techniques to design an advertising system for Twitter. This novel machine learning technique helped to predict customers positive engagement with the ads pushed to their timeline. (McMahan et al. 2013) constructed a deployed CTR prediction system, and combined FTRL-Proximal online learning algorithm and per-coordinate learning rates to improve the accuracy of ad click-through prediction. The difference between our research and these studies is that we focus on the CVR prediction. We utilize different types of features to train a logistic regression model that predicts consumer buying intensions and purchase probability.

Methods

Dataset As have mentioned above, our data is provided by Alibaba for IJCAI-18, detailed descriptions of data are as is shown in the Table 1, 2, 3, 4, and 5.

Data Preprocessing Our training dataset has 478138 transaction records in total. We have given detailed descriptions of variables in the former section. First, we delete all the rows with invalid data cells, which is 368693 records in total. Then, We separate all the data into five tables in order to better discuss them respectively. So the final transaction dataset for us to analyze consists of 109445 records.

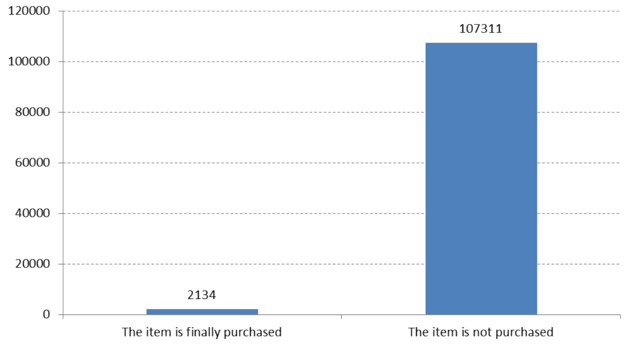


Figure 1: Target value distribution

In Table 1, We can see that instance_id, item_id, user_id, context_id and shop_id can be used as references to search for details of item, user, context and shop information. Is_trade is the target predictive value of our model, since our research focuses on predicting whether there will be a purchase behavior under certain circumstances. In Figure 5, we can see that the proportion of customers who finally purchase is small despite the massive dataset.

Table 1: Clicked Sample

Field	Description
instance_id	Instance ID, of type Long
is_trade	The flag of trade, of type Int. The value of is_trade is 1 if the item is finally purchased, and 0 if not.
item_id	Item ID, of type Long
user_id	User ID, of type Long
context_id	Context ID, of type Long
shop_id	Shop ID, of type Long

In Table 2, for user genders, there are three types of values 0 1 and 2 standing for female, male and family respective; because these three values are equal, so we use one hot encoder to discretize the data, and do the same to user occupations for the same reason. For user age level and user star level, we apply standardization and normalization because they are of different scales.

Table 2: user

Field	Description
user_id	User ID, of type Long
user_gender_id	Gender ID of this user, of type Int. The value is 0 for female users, 1 for male users, and 2 for family users.
user_age_level	User's age level, of type Int. A higher level indicates the user is older.
user_occupation_id	Occupation ID of this user, of type Int
user_star_level	User's star level, of type Int. A higher level indicates a user with more credits.

In Table 3, there are three fields, timestamp, page_id and predict category. Timestamp is in epoch format, so we need to transform it into readable date and time. And we standardize page_id.

Table 3: Context

Field	Description
context_id	Context ID, of type Long
context_timestamp	Time index when this item is displayed, of type Int. The value is a standard Unix timestamp in seconds, and shifted by several days.
context_page_id	ID of the page where the item is displayed, of type Int. The value of the first page is 1, and increases sequentially for the following pages.
predict_category _property	Lists of the predicted categories and properties, of type String.

In Table 4, item price level, sales level, collected level and pv level have been ready for analyzing. Since category list, property list, brand and city are all categorical data, we transform them using one encoder into binary variables.

Table 4: Advertising item

Field	Description
item_id	Item ID, of type Long
item_category_list	Lists of the item's properties, of type String.
item_property_list	Lists of the item's properties, of type String.
item_brand_id	Brand ID of this item, of type Long
item_city_id	City ID of this item, of type Long
item_price_level	Level of price, of type Int and starting from 0. A higher level indicates the item is with higher price.
item_sales_level	Level of sales, of type Int and starting from 0. A higher level indicates the item is purchased for more times.
item_collected_level	Level of collected number, of type Int and starting from 0. A higher level indicates the item is collected for more times.
item_pv_level	Level of displayed number, of type Int and starting from 0. A higher level indicates the item is displayed for more times.

In Table 5, shop review number level, review positive rate, service score, delivery score and description score are all well-format, thus no further processing needed. However, shop star level needs to be discretized.

Table 5: Shop

Field	Description
shop_id	Shop ID, of type Long
shop_review_num_level	Level of review numbers, of type Int and starting from 0.
shop_review_positive_rate	Rate of positive reviews, of type Double. The value is between 0 and 1.
shop_star_level	User's star level, of type Int. A higher level indicates shop with more credits.
shop_score_service	Score of service, of type Double. The value is between 0 and 1.
shop_score_delivery	Score of delivery, of type Double. The value is between 0 and 1
shop_score_description	Score of description, of type Double. The value is between 0 and 1.

Training model Logistic regression models the probability of obtaining a particular value of the nominal variable. In our case, the nominal variables are purchase and no purchase. The dependent variable (Y variable) in logistic regression is the probability that the user finally purchases the advertised item. The probability takes values from 0 to 1. Because the limited range of the probability would present problems if used directly in a regression, so the odds, Y(1-Y) is used instead. Considering the natural log of the odds is more suitable for a regression, the equation of a multiple logistic regression looks like this (eq. 1):

$$Ln[Y/(1-Y)] = \beta_0+\beta_1\times X1+\beta_2\times X2+\beta_3\times X3... \quad (1)$$

The slopes (β_1,β_2,etc) and intercept (β_0) of the best-fitting equation in a multiple logistic regression using the maximum-likelihood method, rather than the least-squares method. The basic idea of Maximum likelihood is to find the values of the parameters under which you would be most likely to get the observed results.

Evaluation Statisticians do not agree on the best measure of fit for multiple function. Some use deviance (D), for which the smaller numbers represent better fit. And some use one of pseudo-R2 values, for which larger numbers represent better fit.

A loss function is the function that maps the value your model produced onto the real value. It indicates magnitude of error your model make on prediction. Because the output of our study is a probability value between 0 and 1, we use a loss function called Cross-Entropy, also known as Log Loss, to calculate the gradient and minimize cost. Cross-Entropy loss increases as the predicted probability diverges from the actual label. A perfect model is to have a log loss of 0.

Cross-entropy loss (eq. 2) can be divided into two separate cost functions: one for y=1 (eq. 3) and one for y=0 (eq. 4).

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^i), y^{(i))}) \quad (2)$$

$$Cost(h_\theta(x), y) = -log(h_\theta(x)) if y = 1 \quad (3)$$

$$Cost(h_\theta(x), y) = -log(1 - h_\theta(x)) if y = 0 \quad (4)$$

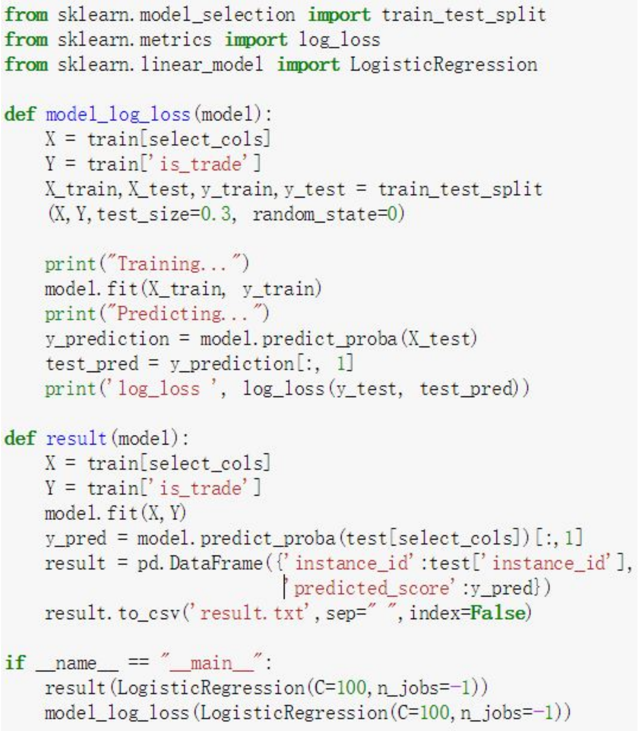
As is shown in eq. 5, the two cost function can be compressed into one.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)}))+(1-y^{(i)})log(1-h_\theta(x^i))] \quad (5)$$

Experimental Design

After data cleaning and data processing using Python pandas library, we start to train a logistic regression model. The variable number increases dramatically after transforming all the categorical data into binary variables. Given that too many features result in overfitting of model, we decide to ignore all the categorical variables to start our first experiment. Therefore, in this experiment, we use 14 variables, which are item price level, item sales level, item collected level, item pv level, user gender id, user age level, user star level, context page id, shop review number level, shop review positive rate, shop star level, shop service score, shop delivery score and shop description score.

We use LogisticRegression package of sklearn library in Python to train and evaluate the model. The code is shown in Figure 2



```
from sklearn.model_selection import train_test_split
from sklearn.metrics import log_loss
from sklearn.linear_model import LogisticRegression

def model_log_loss(model):
    X = train[select_cols]
    Y = train['is_trade']
    X_train, X_test, y_train, y_test = train_test_split
    (X, Y, test_size=0.3, random_state=0)

    print("Training...")
    model.fit(X_train, y_train)
    print("Predicting...")
    y_prediction = model.predict_proba(X_test)
    test_pred = y_prediction[:, 1]
    print('log_loss ', log_loss(y_test, test_pred))

def result(model):
    X = train[select_cols]
    Y = train['is_trade']
    model.fit(X, Y)
    y_pred = model.predict_proba(test[select_cols])[:, 1]
    result = pd.DataFrame({'instance_id':test['instance_id'],
                           'predicted_score':y_pred})
    result.to_csv('result.txt', sep=" ", index=False)

if __name__ == "__main__":
    result(LogisticRegression(C=100, n_jobs=-1))
    model_log_loss(LogisticRegression(C=100, n_jobs=-1))
```

Figure 2: Model training and evaluation in Python

Experimental Results

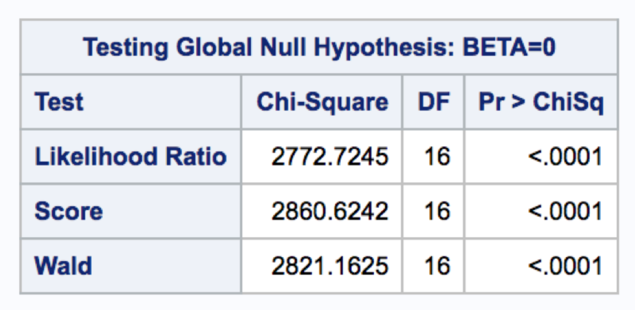
The data set has a binary response variable, which is equal to 1 if a customer purchased the advertised item, and 0 otherwise. We run the logistic regression in SAS. The first part of the output shows that 478138 observations in our data set were used in the analysis and in 9021 observations the customer finished a trade.

The likelihood ratio chi-square of 2772.7245 with a p-value of 0.0001 tells us that our model as a whole fits significantly better than an empty model, as the results of global null hypothesis shown in Figure 3.

According to the type 3 analysis of effects (Figure 4), there are 12 variables Statistically significant at the 5% significance level.

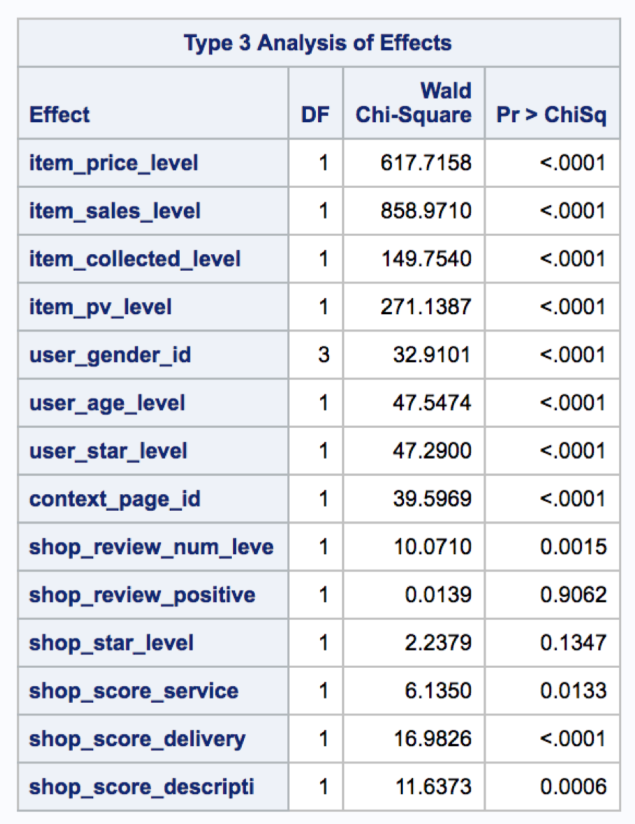
The table of Analysis of Maximum Likelihood Estimate shows the coefficients (labeled Estimate), their standard errors (error), the Wald Chi-Square statistic, and associated p-values. According to the Wald Chi-Square statistic, and associated p-values shown in the table indicates that coefficients for item_price_level, item_sales_level, item_collected_level, item_pv_level, user_age_level, user_star_level, context_page_id, shop_review_num_level, shop_score_service, shop_score_delivery and shop_score_description are statistically significant at a confident level of 5 percent, as are the term of user_gender_id=0 (versus the reference category user_gender_id=1). The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

For a one unit increase in the shop service level, the log odds of purchase increases by 10.9562. For a one unit increase in the shops description score, the log odds of pur-



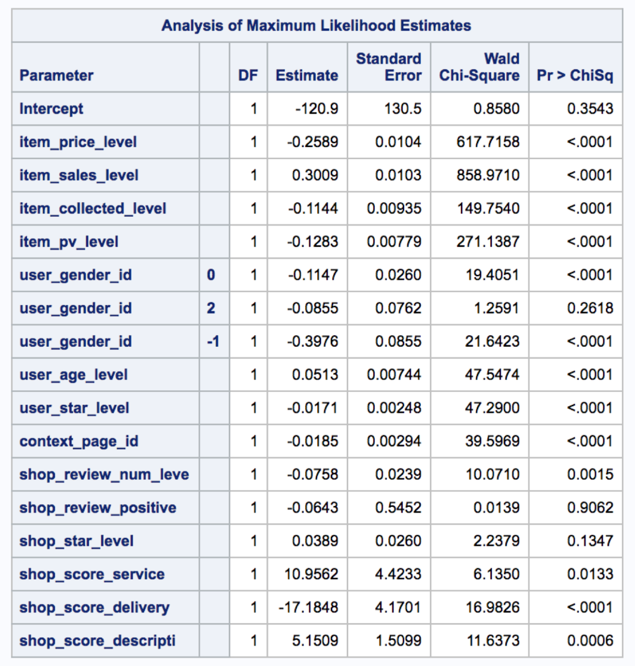
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2772.7245	16	<.0001
Score	2860.6242	16	<.0001
Wald	2821.1625	16	<.0001

Figure 3: Global Null Hypothesis



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
item_price_level	1	617.7158	<.0001
item_sales_level	1	858.9710	<.0001
item_collected_level	1	149.7540	<.0001
item_pv_level	1	271.1387	<.0001
user_gender_id	3	32.9101	<.0001
user_age_level	1	47.5474	<.0001
user_star_level	1	47.2900	<.0001
context_page_id	1	39.5969	<.0001
shop_review_num_leve	1	10.0710	0.0015
shop_review_positive	1	0.0139	0.9062
shop_star_level	1	2.2379	0.1347
shop_score_service	1	6.1350	0.0133
shop_score_delivery	1	16.9826	<.0001
shop_score_descripti	1	11.6373	0.0006

Figure 4: Type 3 Analysis of Effects



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-120.9	130.5	0.8580	0.3543
item_price_level		1	-0.2589	0.0104	617.7158	<.0001
item_sales_level		1	0.3009	0.0103	858.9710	<.0001
item_collected_level		1	-0.1144	0.00935	149.7540	<.0001
item_pv_level		1	-0.1283	0.00779	271.1387	<.0001
user_gender_id	0	1	-0.1147	0.0260	19.4051	<.0001
user_gender_id	2	1	-0.0855	0.0762	1.2591	0.2618
user_gender_id	-1	1	-0.3976	0.0855	21.6423	<.0001
user_age_level		1	0.0513	0.00744	47.5474	<.0001
user_star_level		1	-0.0171	0.00248	47.2900	<.0001
context_page_id		1	-0.0185	0.00294	39.5969	<.0001
shop_review_num_leve		1	-0.0758	0.0239	10.0710	0.0015
shop_review_positive		1	-0.0643	0.5452	0.0139	0.9062
shop_star_level		1	0.0389	0.0260	2.2379	0.1347
shop_score_service		1	10.9562	4.4233	6.1350	0.0133
shop_score_delivery		1	-17.1848	4.1701	16.9826	<.0001
shop_score_descripti		1	5.1509	1.5099	11.6373	0.0006

Figure 5: Maximum Likelihood Estimates

chase increases by 5.1509. For a one unit increase in the sales level of items, the log odds of purchase increases by 0.3009. For a one unit increase in the age level of customers, the log odds of purchase increases by 0.0513.

For every one unit change in the shops delivery score, the log odds of purchase decreases by 17.1848. For every one unit change in the price level of items, the log odds of purchase (versus non-purchase) decreases by 0.2589. For every one unit change in the pv level of items, the log odds of purchase decreases by 0.1283. For every one unit change in the collected level of items, the log odds of purchase decreases by 0.1144. For every one unit change in the level of shop review numbers, the log odds of purchase decreases by 0.0758. For every one unit change in the context page id, the log odds of purchase decreases by 0.0185. For every one unit change in the star level of customers, the log odds of purchase decreases by 0.0171.

With regard to coefficients for the categories of gender, being a female customer, versus being a male customer, decreases the odds of purchase by 0.1147.

Conclusions

In sum, improving a shops service and description will greatly increase the probability of purchase. We recommend shops to provide better service and improve their shop description. They need to adjust the delivery speed to a reasonable level. The results show more times an item is purchased, more possible it will be purchased again. So, shops can also consider improving the CVR from item level, and try to increase the sales level by decreasing the price. The advertisement can target senior and male customers who are more likely to be influenced. Besides, putting the ads on first few pages of the website will increase their exposure, so a higher conversion rate can be achieved.

References

Ansolabehere, S.; Iyengar, S.; Simon, A.; and Valentino, N. 1994. Does attack advertising demobilize the electorate? *American political science review* 88(4):829–838.

Charlton, A., and Blair, V. 1989. Predicting the onset of smoking in boys and girls. *Social science & medicine* 29(7):813–818.

Ciaramita, M.; Murdock, V.; and Plachouras, V. 2008. Online learning from click data for sponsored search. In *Proceedings of the 17th international conference on World Wide Web*, 227–236. ACM.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3(Mar):1157–1182.

Li, C.; Lu, Y.; Mei, Q.; Wang, D.; and Pandey, S. 2015. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1959–1968. ACM.

McMahan, H. B.; Holt, G.; Sculley, D.; Young, M.; Ebner, D.; Grady, J.; Nie, L.; Phillips, T.; Davydov, E.; Golovin, D.; et al. 2013. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1222–1230. ACM.

Merlo, J.; Chaix, B.; Ohlsson, H.; Beckman, A.; Johnell, K.; Hjerpe, P.; Råstam, L.; and Larsen, K. 2006. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology & Community Health* 60(4):290–297.

Richardson, M.; Dominowska, E.; and Ragno, R. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, 521–530. ACM.

Stoltzfus, J. C. 2011. Logistic regression: a brief primer. *Academic Emergency Medicine* 18(10):1099–1104.