Analytics Project

Second Semester, A.Y. 2024-2025

Stat 218 - Statistical Machine Learning

Modeling Religiocentrism in Philippines: Ordinal vs. Binary Logistic Regression

Submitted by:

Alton B. Abejuela

202420043

BACKGROUND OF THE STUDY

Ethnocentrism is a fundamental concept in Psychology and social science. It talks about potential adverse influences on intergroup (in-group) relations and can also lead to hostility towards outgroup (Bizumic and Duckitt, 2012). In earlier studies, it was defined as the technical name in which a person views his own group as the center of everything, and every other groups are scaled and referenced to it (Summer, 1906).

The concept of Religiocentrism revolves around Ethnocentrism, that deals with different religion notions across different ethnicities. It talks about the ingroup's religious values and how the ingroup influences the development of values of an individual towards the outgroup. In the perspective of Social Identity Theory, Religiocentrism were viewed as the difference between members of the same group are seen as peripheral while similarities between members outside the group became central. Members of the outgroup are perceived as having shared ideas and feelings, whereas the characteristics of ingroup are considered more diversified and nuanced (Sterkens and Anthony, 2008).

Previous studies were conducted about ethno-religious groups in the Philippines. A study conducted by Abanes, Kanas, and Scheepers (2014) shows similarity between Christians and Muslims on who they tend to avoid interreligious contact. Both choose to avoid contact with public roles such as mayors and police and on the other hand, they are less likely to avoid private roles who had less power compared to the public roles such as neighbors, classmates and friends. Another study conducted by Camacho, Sterkens, and Scheepers (2017) observed that "While Christian and Muslim youth are not in a permanent stage of manifest conflict, their relationships are nevertheless shaped by mutual prejudices, hostilities and in-group bias".

This study examined the correlation between an individual's religion and level of education to their outlook on why they think that their own religion is the only acceptable religion. While other factors may contribute to this belief, the study focused specifically on the effects of these two variables. This study aims to explain the relationship of religion towards exclusivity beliefs especially for a wide diversity of religion within Philippines. It also aims to study the effect of education in shaping religious exclusivity that can help educators on guiding the society to be more open towards other religion.

The concept of Ethnocentrism is tackled in this study, but this contribution looks into the idea of Religiocentrism,specifically about religion exclusivity, cut across the level of education of the respondents within the Philippines. The focus of this study revolved around the question of "To what extent do individuals believe that their own religion is the only valid religion?".

Methodology

For this study, data from WVS dataset for Philippines (2019) was used to analyze the correlation between the response: The only acceptable religion is my religion (Q170), that is in the Likert scale with 4 levels: Strongly agree (1), Agree (1), Disagree (3), Strongly disagree (4), and the two predictors: Religious Denomination (Q289) and Highest educational level: Respondent (Q275R). The dataset was evaluated into two different models. For the first model, the researcher used ordinal regression to examine the odds ratio between the four levels of the response variable. For the second model, the researcher used Logistic Regression transforming the response variable into a binary type: Agree (1) and Disagree (0). The second model was primarily used to evaluate and compare the performance versus the first model where the four levels were retained.

Many Social Science studies incorporate an ordinal response where Y follows a level or scale, often using 5-point scale (strongly disagree (1), agree (2), neutral (3), agree (4), strongly agree (5)). The first part of this study talks about the application of ordinal logistic regression as the chosen model. With the response variable being ordinal, it is the ideal model that will retain all the information/level for the model. Using metric or nominal model with ordinal data may sometimes lead to over or underestimating the information provided by the data. Fortunately, advancement in statistics and computer science have provided an ideal model for ordinal data. These models are called "ordinal regression models" (Liddell and Kruschke, 2018).

Another method in approaching ordinal response variables is treating it as a binary logistic regression. For example, you can split your data points into those who strongly agree that their religion is the only true religion and those who do not. One downside of it is you would have to do it per each level, creating a series of binary logistic regressions. These binary regressions will replicate the outcomes at different thresholds within our ordinal regression.

The method mentioned is wasteful as multiple binary logistic regressions are created just to compute the outcomes at each level. According to Liddell and Kruschke (2018), another reason why using ordinal regression is better because when the response variable is treated as a metric rather than ordinal, the model is assuming that each level is equidistant. It assumes that the distance between strongly disagree to disagree is equal with the distance with disagree and

agree. Another reason is that ordinal response may not be non-normal as most are distributed to the extremes. Lastly, variances among groups may differ This There are several assumptions when choosing logistic ordinal regression

With ordinal regression, the model focuses on cumulative probabilities rather than the probability at each categorical level. One important assumption with ordinal regression or our cumulative probability model is the Proportional Odds (Parallel lines) Assumption. The interpretation of the proportional odds model is based on the assumption that all observations share the same effect of predictors across the different cumulative splits of the ordinal outcome (Peterson & Harrell, 1990).

The assumption of Proportional Odds is frequently violated, just like the assumptions of normality of residuals or multicollinearity in ordinary regression are frequently violated, regardless the Proportional Odds can still be used as it is still useful. Harrell, F.E. (2015). In this study, a proportional odds model was initially fitted using the ordinal package. When the parallel lines assumption was found to be violated, adjustments were made by applying a partial proportional odds model using the VGAM package to address the issue.

Initially, the full dataset was used for modelling the PO and BLR. Later then a training set and a test set were created (80% training set / 20% test set) to test out the accuracy of the created model just to observe possible overfitting and underfitting of models.

Partial Proportional Odds (PPO) model

Looking back at the distribution of the response variable per Religious denomination (Table 1.2.1), some groups received a very small frequency across different levels. Since the violation of Proportional Odds assumption was only seen with one of the variables, grouping of these smaller denominations can be done to proceed with the PPO model with a tradeoff that most of the information about the religion will be lost since the comparison for the Religion will basically become Roman Catholic vs Other Religions.

The researcher chose not to proceed with this model as it can be clearly seen that there was an issue with the number of the dataset so the validity of the previous tests for Proportional Odds assumption might be less variable. It should be noted that limitations with the number of observations in some groups may affect the reliability of the tests for the PO assumption (Peterson & Harrell, 1990).

Binary Logistic Regression Model

For this model, the ordinal response variable was transformed into a binary response to apply a logistic regression. The four levels of the *Y* were divided into two groups– Disagree (0) which combined Strongly disagree and Disagree; Agree (1), which combined Strongly agree and Agree. It was assumed that the relationship between the independent variables and the logit of the dependent variable was linear, and that there were no interaction effects among the covariates.

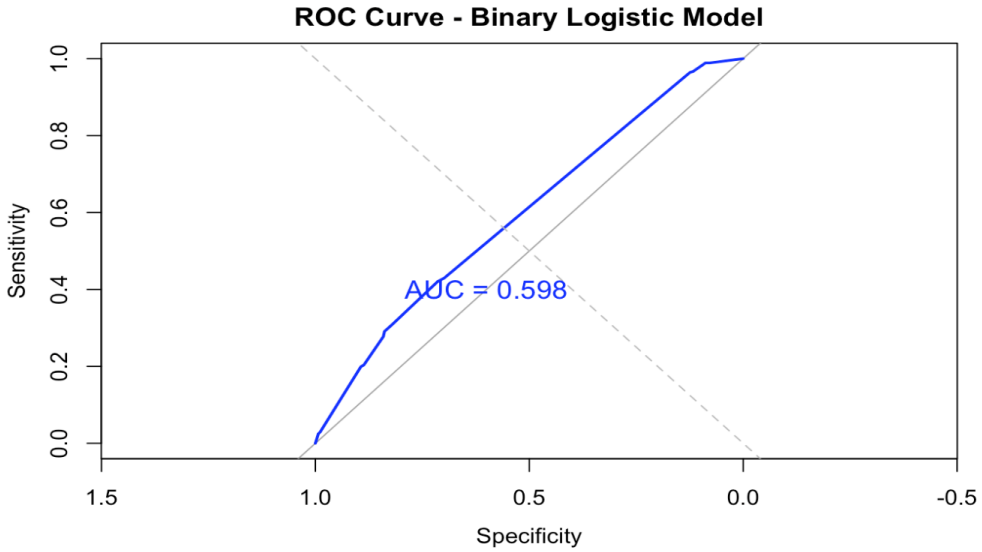

Figure 3.1.1 ROC curve for binary logistic model

The model hows an AUC of 0.598 - this indicates that the model predicts slightly better than random chance

Results and Discussion

Exploratory Data Analysis

For this study, data from the 2019 World Value Surveys Association (WVS) for Philippines was used. The dataset will include question "The only acceptable religion is my religion?" (Q170) as the response variable with two predictors - Religious denominations - major groups (Q289) and Highest educational level: Respondent (recoded into 4 groups) (Q275R). Values with either Don't know (-1), No answer (-2), Not applicable (-3), Not asked (-4) and Missing; Not available (-5) are excluded for easier analysis.

Table 1.1.1 Frequency distribution of Response Variable - *"The only acceptable religion is my religion?"*

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Strongly agree	261	22.29	22.29	22.29	22.29
Agree	463	39.54	61.83	39.54	61.83
Disagree	389	33.22	95.05	33.22	95.05
Strongly disagree	58	4.95	100.00	4.95	100.00
<NA>	0			0.00	100.00
Total	1171	100.00	100.00	100.00	100.00

Overall respondents voted 22.29% (261) Strongly agree, 39.54% (463) for Agree, 33.22% (389) for Disagree, 4.95% (58) for Strongly Agree. For the cumulative frequency 22.29% entered the top level, 61.83% entered the upper level, 95.05% entered the lower level and 100% for the bottom level. Strongly agree will be the highest level down to Strongly disagree as the lowest level.

The relationship of the response variable with Religion and Education level was examined. First, Table 1.2.1 shows the distribution of the response variable with Religion of the respondent and Table 1.3.1 shows the distribution of the response variable with Level of education of the respondent.

Table 1.2.1 Response variable by Religious Denomination

Q170 religion	Strongly agree	Agree	Disagree	Strongly disagree	Total
Roman Catholic	162 (17.5%)	400 (43.1%)	318 (34.3%)	47 (5.1%)	927 (100.0%)
Muslim	57 (76.0%)	11 (14.7%)	6 (8.0%)	1 (1.3%)	75 (100.0%)
Other	11 (32.4%)	7 (20.6%)	15 (44.1%)	1 (2.9%)	34 (100.0%)
Other Christian	17 (37.0%)	16 (34.8%)	12 (26.1%)	1 (2.2%)	46 (100.0%)
Protestant	14 (15.7%)	29 (32.6%)	38 (42.7%)	8 (9.0%)	89 (100.0%)
Total	261 (22.3%)	463 (39.5%)	389 (33.2%)	58 (5.0%)	1171 (100.0%)

Most of the respondents belong to the Roman Catholic group with Protestant and Muslim next to it. Looking at the table above, Although Roman Catholic has the highest number of respondents who entered the top level, looking at its proportion towards its own group, 17.5% only entered the top level compared to the Muslim group where 76% of its group proportion

entered the top level. For the upper level, there's a high number and proportion of Roman Catholic who entered this level. Overall, a total of 61.8% entered the top 2 levels.

This can be attributed to the distribution of religion within the Philippines where according to PSA (2020), 79% of the population is primarily Christian, with Roman Catholicism being the dominant faith and Islam were shown as among the minorities (6.4%).

Table 1.3.1 Response variable by Education Level

Q170	Strongly agree	Agree	Disagree	Strongly disagree	Total
education					
Primary	193 (23.3%)	343 (41.4%)	256 (30.9%)	37 (4.5%)	829 (100.0%)
Secondary	33 (21.3%)	63 (40.6%)	52 (33.5%)	7 (4.5%)	155 (100.0%)
Post-secondary	35 (18.7%)	57 (30.5%)	81 (43.3%)	14 (7.5%)	187 (100.0%)
Tertiary	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total	261 (22.3%)	463 (39.5%)	389 (33.2%)	58 (5.0%)	1171 (100.0%)

Looking at the Education Level, those with the Primary as the highest level of education have the highest frequency. Secondary and Post-secondary have almost the same count while none of the respondents has Tertiary as their highest educational attainment. Primary, Seconday and Post-secondary have the similar proportion to those who entered the top level. For the upper level, Primary and Secondary have similar proportions while Tertiary has the lowest (30.5%). Overall, 61.9% of the respondents agree with the response variable.

Proportional Odds (PO) model

For the first model of this study, Proportional Odds model was used from the ordinal package with link as logit. It was first assumed that all assumptions about PO are met to examine its performance compared to other models. After fitting the model, different tests were conducted to test the assumptions for the Proportional Odds model. It is important to note that the result of these tests can be greatly affected by the number of overactions.

Table 2.2.1 Goodness of fit using Likelihood Ratio Test

	no.par	AIC	logLik	LR.stat	df	Pr(>Chisq)	
model_po_null	3	2854.8	-1424.4				
model_po	9	2740.4	-1361.2	126.41	6	< 2.2e-16	***

Likelihood Ratio Test (LRT) comparing the full model vs the null model indicated a significantly good fit where $P < 0.01$. The result suggests that the two predictors included in the model explain a significant amount of variance within the response variable.

Table 2.3.1 Proportional Odds Assumption using Nominal Test

	Df	logLik	AIC	LRT	Pr(>Chi)	
<none>		-1361.2	2740.4			
education	4	-1359.0	2744.0	4.3469	0.36109	
religion	8	-1353.0	2740.0	16.3315	0.03787	*

Table 2.3.2 Brant test

Test for	X2	df	probability
Omnibus	25.9	12	0.01
education.L	4.09	2	0.13
education.Q	0.53	2	0.77
religionMuslim	7.14	2	0.03
religionOther	13.7	2	0
religionOther Christian 1.73	2	0.42	
religionProtestant	1.81	2	0.4

Looking at the nominal and brant test, the assumption of proportional odds was violated, specifically by the Religious Denomination variable. Referring to the study of Harrell, F.E. (2015), the model can still mean something as PO assumption is usually violated like the assumptions on a normal regression model because during the evaluation of the coefficients of each variable, both predictors shown a significant effect with the response variable (based on each p-value). These score-based test can be oversensitive showing significant violations even when there might not be any real problem.

Table 4.1.1 PO model summary

link <chr>	threshold <chr>	nobs <dbl>	logLik <chr>	AIC <chr>	niter <chr>	max.grad <chr>	cond.H <chr>
logit	flexible	1171	-1361.18	2740.37	6(0)	1.81e-11	4.9e+01

To examine the effects of the two predictors: Religious denomination and Level of Education of the respondent with the response variable, two different models were conducted by the researcher. For the first model, ordinal regression, specifically the Proportional Odds model

was applied, and it was initially assumed the PO assumption was never violated. After several evaluation It was discovered that the PO assumption was violated by one of the predictors – Religion. The researcher attempted to conduct Partial Proportional Odds model towards the dataset to address this issue, allowing for a linear effect of education across response categories while relaxing the assumption for religion to allow a non-proportional (non-parallel) effect. Upon analysis, it was discovered that PPO model cannot be applied directly to the dataset without transforming the grouping of the religion. Due to this, the PO model was still used as the chosen model for Ordinal regression in order to retain the original groupings with all the predictors. Moving forward, binary logistic regression model was also conducted with dichotomized version of the response variable (Agree = 1 vs Disagree = 0).

Table 4.1.1 Performance Comparison Table

Model <chr>	AIC <dbl>	BIC <dbl>	Loglik <dbl>	McFadden_R2 <dbl>	CoxSnell_R2 <dbl>	Nagelkerke_R2 <dbl>	Train_Accuracy <dbl>	Test_Accuracy <dbl>
Proportional Odds	2740.366	2785.957	-1361.1832	0.04437310	0.10232717	0.11217540	0.4685	0.4188
Binary Logistic	1514.463	1549.922	-750.2316	0.03643587	0.04729749	0.06430911	0.6233	0.6197

In terms of the overall model fit, the Binary Logistic model shows better performance based on significantly lower AIC and BIC values and a higher log-likelihood, indicating that it describes the observed data more efficiently.

However, looking at the pseudo- R^2, the PO model yields a slightly higher McFadden's R^2, significantly higher CoxSnell and Nagelkerke suggesting that relative to a null model, it might have explained the variation more compared to the Binary Logistic model.

In terms of predictive performance, the Binary Logistic model performs significantly better compared to the PO model. Both models show a similar accuracy for training and test set, this might indicate a balance trade-off between its bias and variance since neither model suffers from overfitting or underfitting.

In conclusion the Binary Logistic Regression model seems to be a better fit for the model and its predictive accuracy is higher compared to the PO model, making it a more suitable choice for classification performance. However, the response variable was transformed into a binary so some of the data between the extremes were lost. This might explain why the variability of the dataset were explained better by the PO model since the ordinal structure was retained. Since this study aims to explain the effect of level of education towards Religiocentrism, the PO model is a better model when it comes to explaining the effects of the predictors towards the response variable in a way where none of the information were lost.

Ordinal data that are analyzed using metric models assumes that the differences between each interval are equal, which is not true, this can often lead to misinterpretations of the data. Treating ordinal data with metric models can produce type 1 and type 2 errors which can also produce *inversion of effects*. Liddell and Kruschke (2018). This can be later seen when comparing the estimates of predictors from PO model compared to BLR.

Interpretation for the Proportional Odd Model

Looking at Table 4.2.1, it can be observed that the odds of entering a higher-level decrease by approximately 30% (exponent of the coefficient -0.36, $\exp(-0.36) \approx 0.70$) for each unit increase in education (linear effect). This suggests that education might make the respondents more open towards out-group religion, and they are less certain about the idea that their own religion is the only valid one. According to the study of Sterkens and Anthony (2008), higher level of education shows greater tolerance towards religious outgroups and a strong predictor for lower level of religiocentrism.

Another study of Meiza (2019) about Muslim Indonesian students shows a relevant relationship between a person's tolerance level, how open they are toward religious outgroups, and their radical intentions. Radical intentions include early stages — pre-radical attitudes, such as religious exclusivity, where individuals begin to see their own religion as the only valid one and reject others. It shows how formal education plays a role on shaping an individual's openness to other religion.

For the Odds Ratio of the religion, the odds of entering a higher level increase by approximately 1149% (exponent of the coefficient 2.52, $\exp(2.52) \approx 12.49$) for individuals identifying as Muslim compared to the reference religion, Roman Catholic. This means that Muslims in the study are more convinced that their religion own religion is the only true religion compared to the out-group religion. On the other hand, Protestants tend to be less likely to strongly agree with that idea than Roman Catholics

Other religious groups such as "Other Christian" also show a higher tendency to agree to that idea compared to Roman Catholics, while some groups like "Other" do not really show a significant effect towards that idea. This ideology can be greatly affected not only by the teachings within their own religion but also by the interactions of an individual towards another group. Referencing the study of Sterkens and Anthony (2008), Muslims shows the strongest positive ingroup attitudes compared to Christians and Hindus, this suggests a strong pride towards their own religious beliefs and community. This can be related to their own religious

teachings and other cultural and socio-religious factors. It can also be seen that Christian, in general shows a stronger negative outgroup attitude towards the two religions however, Muslims shows stronger negative towards Christians than Hindus. This is linked into their historical relationship and other political and religious conflicts.

Table 4.2.1 Parameter Estimates for the Proportional Odds Model

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
education.L             -0.3559     0.1092  -3.258  0.00112 **
education.Q             -0.1358     0.1372  -0.990  0.32233
religionMuslim           2.5247     0.2821   8.951  < 2e-16 ***
religionOther            0.1548     0.3439   0.450  0.65262
religionOther Christian  0.8585     0.2909   2.951  0.00317 **
religionProtestant      -0.4268     0.2097  -2.035  0.04180 *
```

Table 4.2.2 Threshold Coefficients for the Proportional Odds Model

```
                           Estimate Std. Error z value
Strongly disagree|Disagree -2.81595    0.14354 -19.618
Disagree|Agree             -0.28904    0.07983  -3.621
Agree|Strongly agree        1.58865    0.09436  16.836
```

Interpretation for Binary Logistic Regression model (For comparison)

Looking at Table 4.3.1, it can be observed that the odds of agreeing increase by approximately 56% (exponent of the coefficient 0.44, exp(0.44) ≈ 1.56) for each unit increase in education (linear effect). This suggests that, in contrast with the Proportional Odds model where education was associated with lower odds of stronger agreement, the Binary Logistic Regression model shows that higher education is associated with a stronger belief that their own religion is the only valid one.

For the Odds Ratio of the religion, the odds of agreeing decrease by approximately 83.5% (exponent of the coefficient -1.80, exp(-1.80) ≈ 0.17) for individuals identifying as Muslim compared to the reference religion, Roman Catholic. Respondents who are Muslim show a dramatically lower likelihood of agreeing with the statement compared to Roman Catholics, this result directly contradicts the findings in the PO model, where Muslims had much higher odds of stronger agreement compared to Roman Catholics.

The contrast between the two models shows that the relationship between education and Religiocentrism is complex and depending on how the data was categorized, the result can be greatly affected. As demonstrated by Liddell and Kruschke (2018), BLR may yield inverted effects, where the direction of estimated coefficients does not reflect the true ordering of response.

Table 4.3.1 Parameter Estimates for the Binary Logistic Regression model

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.30564    0.08368  -3.652  0.00026 ***
education.L               0.44193    0.11797   3.746  0.00018 ***
education.Q               0.17546    0.15298   1.147  0.25138
religionMuslim           -1.79898    0.40346  -4.459 8.24e-06 ***
religionOther             0.31401    0.35259   0.891  0.37315
religionOther Christian  -0.61874    0.33910  -1.825  0.06805 .
religionProtestant        0.49663    0.22428   2.214  0.02680 *
```

Suggestions for future studies

Future studies are encouraged to utilize a bigger and diverse dataset for future analysis. It can help when modeling for Partial Proportional Odds model since it can address the issue of sparsity for each predictor at each level. This will also help with the reliability of assumption tests for the Proportional Odds model, such as the Brant test and the nominal test, by providing more stable estimates and increasing the validity of the results.

Optimization with the model can be also done to improve its results such as different link functions, different cross-validation methods or exploration of different independent variables can be tested to see what other factors can affect an individual's ideology of religious exclusivity.

References

Abanes, M. S., Kanas, A., & Scheepers, P. L. H. (2014). Ethno-religious groups, identification, trust and social distance in the ethno-religiously stratified Philippines. Research in Social Stratification and Mobility, 37, 61–75. https://doi.org/10.1016/j.rssm.2014.02.001

Bizumic, B., & Duckitt, J. (2012). What is and is not ethnocentrism? A conceptual analysis and political implications.

Camacho, A., Sterkens, C., & Scheepers, P. L. H. (2017). Ethno-religious identification and latent conflict: Support of violence among Muslim and Christian Filipino children and youth. In A. Pandya & E. Laipson (Eds.), Islam and politics: Renewal and resistance in the Muslim World (pp. 17–31). The Henry L. Stimson Center. https://doi.org/10.1007/978-981-4585-98-9_12-1

Harrell, F.E. (2015). Ordinal Logistic Regression. In: Regression Modeling Strategies. Springer Series in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-19425-7_13

Liddell, T. and Kruschke, J. (2018), Analyzing ordinal data with metric models: What could possibly go wrong?,Journal of Experimental Social Psychology,Volume 79,Pages 328-348,ISSN 0022-1031,https://doi.org/10.1016/j.jesp.2018.08.009.

Meiza, A. (2023). The ordinal regression to analyze radical intention of Muslim Indonesian students through personality type and tolerance approach. Pakistan Journal of Statistics and Operation Research, 19(2), 359–368. https://doi.org/10.18187/pjsor.v19i2.3932

Peterson, B., & Harrell, F. E., Jr. (1990). Partial proportional odds models for ordinal response variables. Journal of the Royal Statistical Society: Series C (Applied Statistics), 39(2), 205–217. https://doi.org/10.2307/2347760

Summer (1906). Folkways – A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals.

Sterkens, C. and Anthony, F.-V. (2008). A Comparative Study of Religiocentrism among Christian, Muslim and Hindu Students in Tamil Nadu, India.

Appendices

Confusion matrix for PO model (training and test set)

```
[1] "Training Set"

                   Strongly disagree Disagree Agree Strongly agree
  Strongly disagree                 0        0     0              0
  Disagree                         17       87    53             29
  Agree                            26      224   312            135
  Strongly agree                    0        4    10             40
[1] "Test Set"

                   Strongly disagree Disagree Agree Strongly agree
  Strongly disagree                 0        0     0              0
  Disagree                          1       19    25              8
  Agree                            13       53    62             32
  Strongly agree                    1        2     1             17
```

Confusion matrix for BLR model (training and test set)

```
[1] "Training Set"

pred_train_binary   0   1
                0 574 348
                1   5  10
[1] "Test Set"

pred_test_binary   0   1
               0 142  86
               1   3   3
```