Final assignment group 4

Deepika Dilip, Tora Mullings, Daniel Sullivan, Deepa Sharma, Bikram Barua, Newman Okereafor

2022-11-24

Contents

Abstract:

Maternal mortality is a leading public health issue in Bangladesh, with 173 deaths per 100k births. Yet with improvements in public health surveillance, a preventative responsive could be better informed with biomarker data and accurate risk predictions. For this project, we utilize multinomial models to quantify the contribution of biomarkers in predicting mortality risk.

Key words:

maternal health, clinical outcomes

Introduction:

Maternal mortality is a leading public health issue in Bangladesh. Advances in public health outreach and medical pipelines have reduced maternal mortality rates, but there remains a glaring gap, especially when considering additional factors, such as socioeconomic status. One of the WHO Sustainable Development Goals was to reduce the global mortality ratio to less than 70 deaths per 100k births

Here, we further explore mortality risk as a product of standard clinical indicators. We obtained this dataset from the UCI repository. Data was aggregated from different sites, including rural and urban health centers.

According to the WHO approximately 810 women die daily due to pregnancy complications (1). With such a high rate of death associated with childbirth it is important to maximize early interventions in high-risk pregnancies in order to monitor and start early intervention to save both the lives of the mother and child. Because of this need, pregnancy has been the focus of many data scientists research in developing many predictive algorithms to try and aid in identifying at risk pregnancies, best emergency interventions and various other aspects to help both mothers and doctors. For this reason, we want to look at identifying low mid and high-risk pregnancies through regression and machine learning methods in order to aid in identifying individuals who could be helped through early intervention.

Literature review:

At risk pregnancies remain a hot topic of research despite many advances in technology and a shrinking pregnancy/childbirth mortality rate. Predictive modeling has been implemented in several ways to aid in reducing pregnancy risks. There are three major groups of studies that have been performed. There were three major areas of research in these studies. The largest group predicted risks and complications involved with the pregnancy in specific scenarios (3)() as we are trying to asses with our data set. Many papers also covered predicting delivery methods as well as successful vaginal delivery (2). And the last big area of study looks at predicting in vitro fertilization success rates. Our analysis Is of the first group where we are trying to predict at risk pregnancies however the area where we differ from most of these studies is through scope. Most studies that are trying to predict complications do it in a much more specific scope. For example, some studies only predict preterm birth, or complications with vaginal birth while our approach just focuses on a generally high-risk birth and works off mostly basic vitals. Additionally, our analysis works through generalized linear models and progresses into simpler machine learning where as these other studies implement more in depth and domain specific methods.

1 Trends in maternal mortality 2000 to 2017: estimates by who, unicef, unfpa, world bank group and the united nations population division. https://www.unfpa.org/featured-publication/trends-maternal-mortality-2000-2017. Accessed 10 Jan 2021.

2 Birara M, Gebrehiwot Y. Factors associated with success of vaginal birth after one caesarean section (vbac) at three teaching hospitals in addis ababa, ethiopia: a case control study. BMC Pregnancy Childbirth. 2013;13(1):1–6.

3 Gao C, Osmundson S, Edwards DRV, Jackson GP, Malin BA, Chen Y. Deep learning predicts extreme preterm birth from electronic health records. J Biomed Inform. 2019;100:103334.

4 Islam, Muhammed N, Mustafina, Sumaiya N, Mahmud, Tahsin, Khan, Nafiz I Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda

Methodology:

Exploratory Data Analysis:

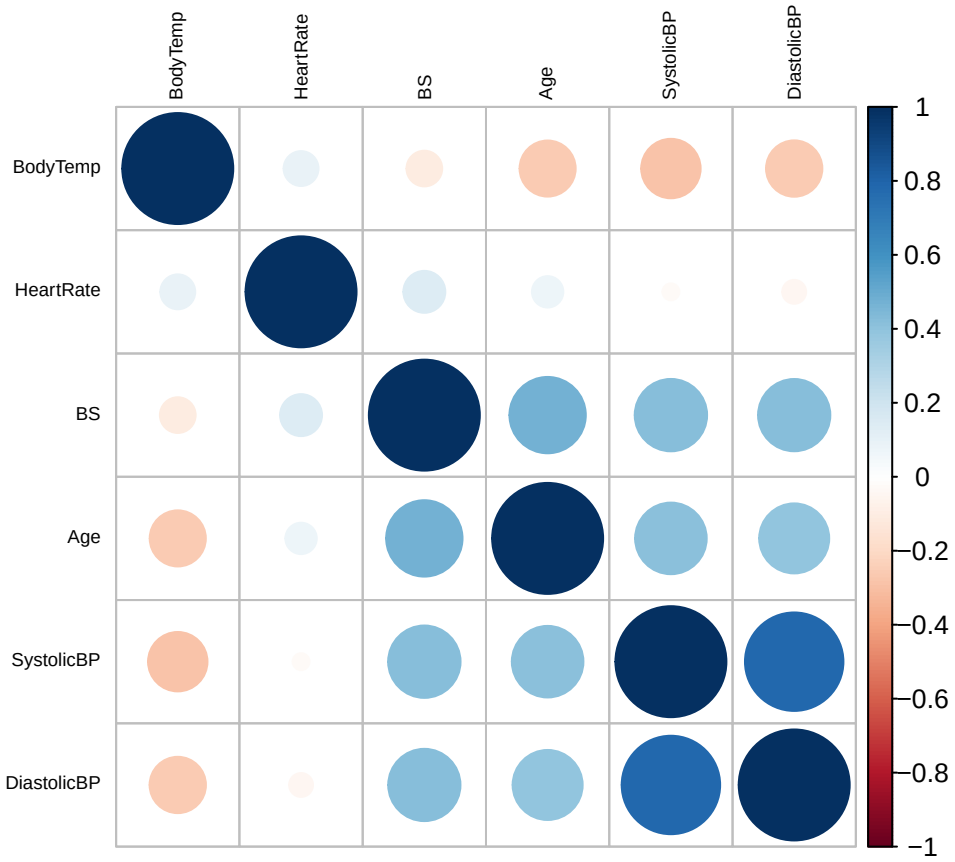
Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
25	130	80	15.0	98	86	high risk
35	140	90	13.0	98	70	high risk
29	90	70	8.0	100	80	high risk
30	140	85	7.0	98	70	high risk
35	120	60	6.1	98	76	low risk

Data Attributes

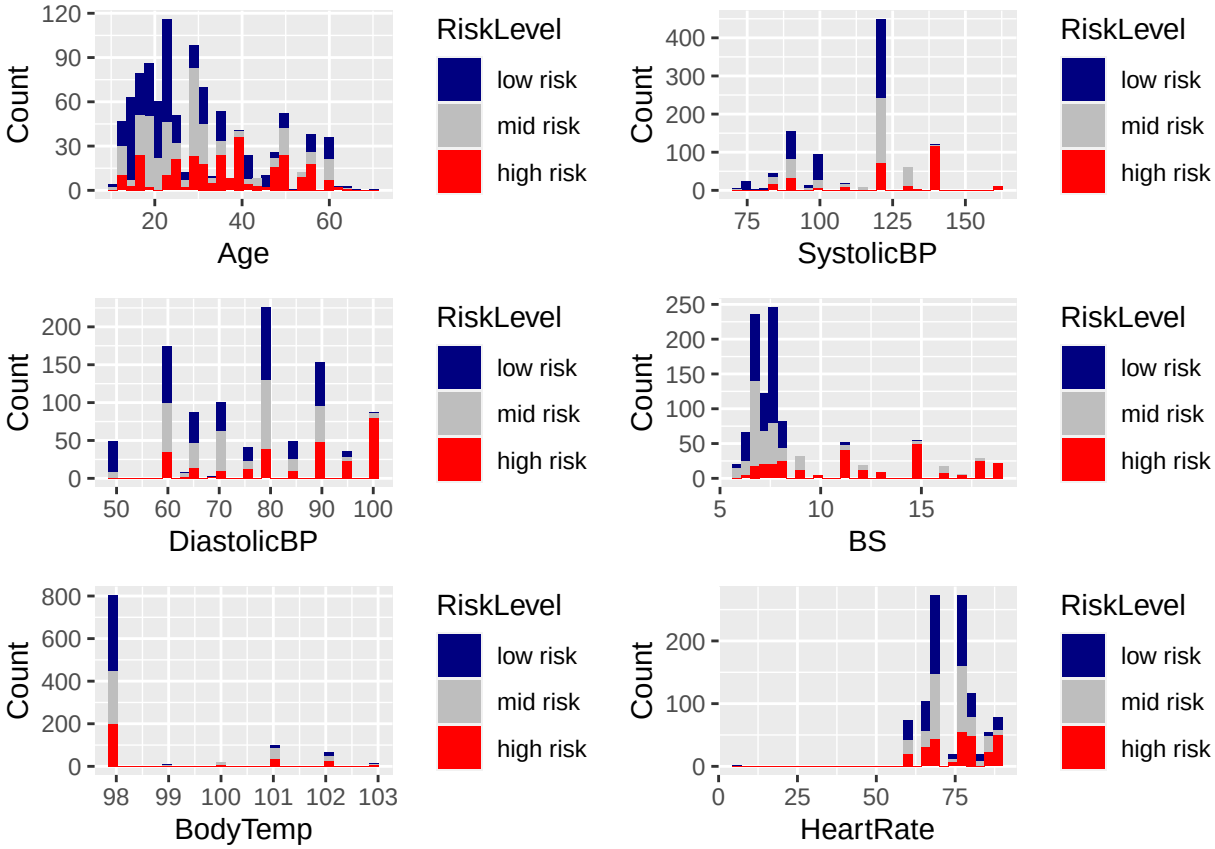
- `Age`: Any ages in years when a women during pregnant.
- `SystolicBP`: Upper value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- `DiastolicBP`: Lower value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- `BS`: Blood glucose levels is in terms of a molar concentration, mmol/L.
- `HeartRate`: A normal resting heart rate in beats per minute.
- `Risk Level`: Predicted Risk Intensity Level during pregnancy considering the previous attribute.

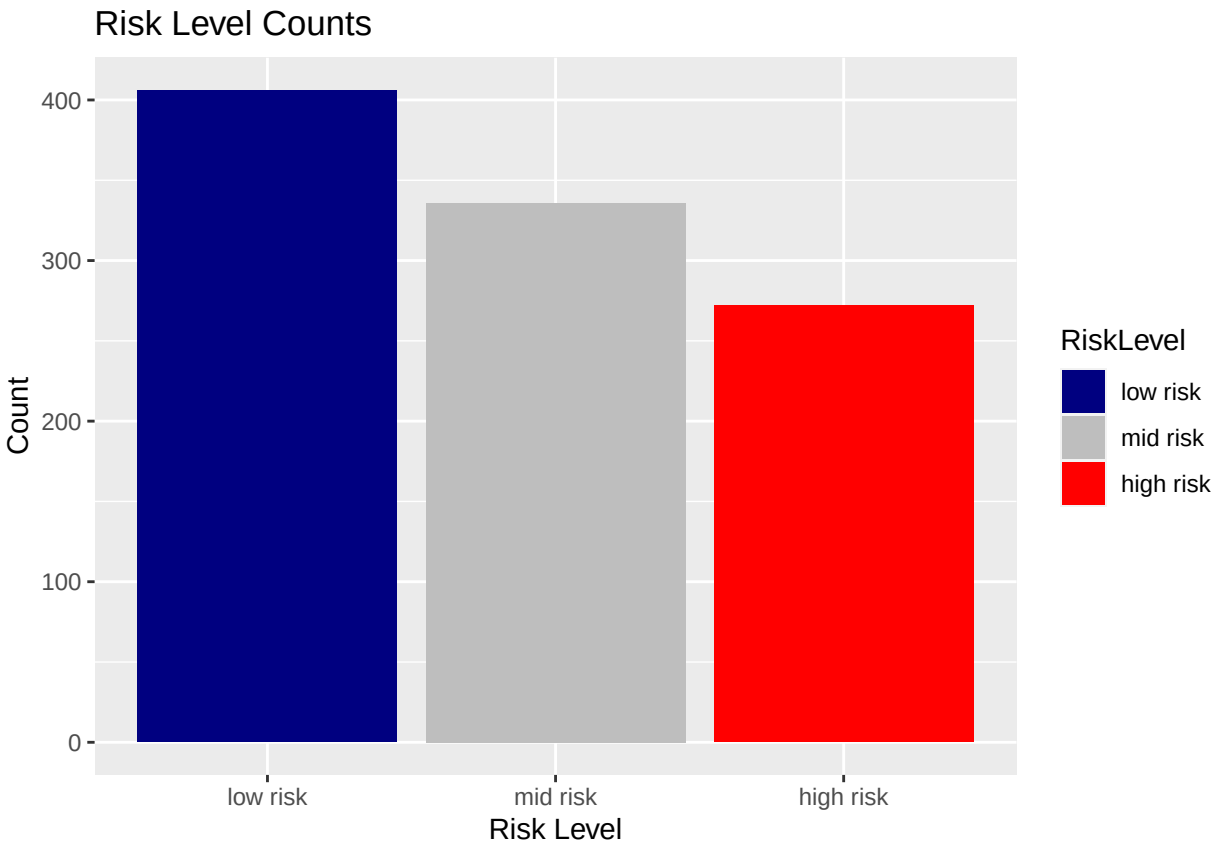
Data exploration

We can start by making a correlation plot to compare continuous values. Age is positively correlated with systolic and diastolic blood pressure.



The first step is visualizing data distribution by risk. Here we can see age is skewed, and the extremities of blood pressure and glucose levels are flagged as high risk.


Count
Age
SystolicBP
DiastolicBP
BS
BodyTemp
HeartRate
RiskLevel
low risk
mid risk
high risk



One approach we can take is implementing unsupervised clustering since many of the biomarkers are continuous. We can do this by forming a matrix of indicators and seeing if risk levels clusters. From this analysis, we see 6 distinct subgroups: 2 high risk, 3 low risk, and 1 mid risk (with some heterogeneity).

Multinomial Regression

First we partitioned the dataset using a 70-30 split. We initially fit a full model with all included variables as predictors. Next, we fit a series of multinomial models, starting with a full model. We then implemented feature selection based on statistical significance to improve accuracy.

Experimentation and Results:

Describe the specifics of what you did (data exploration, data preparation, model building, model selection, model evaluation, etc.), and what you found out (statistical analyses, interpretation and discussion of the results, etc.).

XGBoost

We also decided to try using a model that combined previous models with new ones, subsequently increasing accuracy. Therefore, we decided to fit the eXtreme Gradient Boosting algorthim from the `xgboost` package. In this case, however, we have to split our outcome: one model with predict high risk while the other will predict medium risk.

	X.Intercept.	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
mid risk	-215183.7	-1.226472	6.925812	-2.8623659	3.744558	26.07746	2.753338
high risk	-663191.8	-2.020653	6.769766	0.8518392	7.845787	45.90702	3.926987

Results

Full Model:

Resulting Coefficients:

Confusion Matrix:

```
##
## predicted_class low risk mid risk high risk
##       low risk        82       42         6
##       mid risk        41       39        17
##       high risk        6       15        57
```

```
## # weights:  24 (14 variable)
## initial  value 778.916113
## iter  10 value 641.429163
## iter  20 value 574.691352
## iter  30 value 541.544784
## iter  40 value 541.461488
## iter  50 value 540.621356
## final  value 540.613392
## converged
```

```
##
## predicted_class low risk mid risk high risk
##       low risk        82       42         6
##       mid risk        41       39        17
##       high risk        6       15        57
```

```
## [1] "accuracy=0.583606557377049"
```

```
## [1] "AIC=1109.22678364781"
```

	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.6356589	0.4062500	0.7125000
Specificity	0.7272727	0.7224880	0.9066667
Pos Pred Value	0.6307692	0.4020619	0.7307692
Neg Pred Value	0.7314286	0.7259615	0.8986784
Precision	0.6307692	0.4020619	0.7307692
Recall	0.6356589	0.4062500	0.7125000
F1	0.6332046	0.4041451	0.7215190
Prevalence	0.4229508	0.3147541	0.2622951
Detection Rate	0.2688525	0.1278689	0.1868852
Detection Prevalence	0.4262295	0.3180328	0.2557377
Balanced Accuracy	0.6814658	0.5643690	0.8095833

Age and Systolic BP as Predictors:

Model Coefficients

	X.Intercept.	Age	SystolicBP	BS
mid risk	-6.400819	-1.678529	5.233049	4.115945
high risk	-10.357928	-2.886336	6.661841	8.457483

Confusion Matrix

```
##
## predicted_class2 low risk mid risk high risk
##        low risk        85       49        16
##        mid risk        42       37        15
##        high risk        2       10        49

## # weights:  15 (8 variable)
## initial  value 778.916113
## iter  10 value 605.310909
## final  value 596.349922
## converged
```

	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.6589147	0.3854167	0.6125000
Specificity	0.6306818	0.7272727	0.9466667
Pos Pred Value	0.5666667	0.3936170	0.8032787
Neg Pred Value	0.7161290	0.7203791	0.8729508
Precision	0.5666667	0.3936170	0.8032787
Recall	0.6589147	0.3854167	0.6125000
F1	0.6093190	0.3894737	0.6950355
Prevalence	0.4229508	0.3147541	0.2622951
Detection Rate	0.2786885	0.1213115	0.1606557
Detection Prevalence	0.4918033	0.3081967	0.2000000
Balanced Accuracy	0.6447983	0.5563447	0.7795833

Blood Sugar and Systolic BP as Predictors:

Model Coefficients:

	X.Intercept.	BS	SystolicBP
mid risk	-6.267652	3.899811	5.007696
high risk	-10.173196	8.065761	6.146324

Confusion Matrix:

```
##
## predicted_class3 low risk mid risk high risk
##        low risk        91       65        16
##        mid risk        36       21        16
##        high risk        2       10        48
```

```
## [1] "accuracy=0.560655737704918"
```

```
## [1] "AIC=1208.6998445842"
```

multinomial model 3 (blood sugar, systolic blood pressure)

```
## # weights:  12 (6 variable)
## initial  value 778.916113
## iter  10 value 602.107553
## final  value 601.082024
## converged
```

```
## [1] 1214.164
```

	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.7054264	0.2187500	0.6000000
Specificity	0.5397727	0.7511962	0.9466667
Pos Pred Value	0.5290698	0.2876712	0.8000000
Neg Pred Value	0.7142857	0.6767241	0.8693878
Precision	0.5290698	0.2876712	0.8000000
Recall	0.7054264	0.2187500	0.6000000
F1	0.6046512	0.2485207	0.6857143
Prevalence	0.4229508	0.3147541	0.2622951
Detection Rate	0.2983607	0.0688525	0.1573770
Detection Prevalence	0.5639344	0.2393443	0.1967213
Balanced Accuracy	0.6225995	0.4849731	0.7733333

Blood Sugar as Predictor:

Coefficients:

	X.Intercept.	BS
mid risk	-4.154068	3.919225
high risk	-10.340427	8.994715

```
## [1] "accuracy=0.524590163934426"
```

```
## [1] "AIC=1214.16404899514"
```

multinomial model 4 (blood sugar)

```
## # weights:  9 (4 variable)
## initial  value 778.916113
## iter  10 value 626.776588
## final  value 626.754462
## converged
```

Confusion Matrix:

```
##
## predicted_class2 low risk mid risk high risk
##        low risk        85       49        16
##        mid risk        42       37        15
##        high risk        2       10        49
```

	X.Intercept.	BS
mid risk	-4.154068	3.919225
high risk	-10.340427	8.994715

	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.9689922	0.0520833	0.6500000
Specificity	0.4204545	0.9569378	0.9466667
Pos Pred Value	0.5506608	0.3571429	0.8125000
Neg Pred Value	0.9487179	0.6872852	0.8838174
Precision	0.5506608	0.3571429	0.8125000
Recall	0.9689922	0.0520833	0.6500000
F1	0.7022472	0.0909091	0.7222222
Prevalence	0.4229508	0.3147541	0.2622951
Detection Rate	0.4098361	0.0163934	0.1704918
Detection Prevalence	0.7442623	0.0459016	0.2098361
Balanced Accuracy	0.6947234	0.5045106	0.7983333

XGBoost Model

Predicting High Risk

	x
Sensitivity	0.9733333
Specificity	0.9875000
Pos Pred Value	0.9954545
Neg Pred Value	0.9294118
Precision	0.9954545
Recall	0.9733333
F1	0.9842697
Prevalence	0.7377049
Detection Rate	0.7180328
Detection Prevalence	0.7213115
Balanced Accuracy	0.9804167

```
## [1] "accuracy=0.59672131147541"
```

```
## [1] "AIC=1261.50892465212"
```

multinomial model 5 (blood sugar, heart rate, body temp, systolic blood pressure)

```
## # weights:  18 (10 variable)
## initial  value 778.916113
## iter  10 value 611.387656
```

```
## iter  20 value 552.827818
## iter  30 value 550.983997
## iter  40 value 550.267068
## final  value 550.220311
## converged
```

```
##
## predicted_class5 low risk mid risk high risk
##        low risk        87       42         4
##        mid risk        36       37        24
##        high risk        6       17        52
```

```
## [1] "accuracy=0.577049180327869"
```

```
## [1] "AIC=1120.44062122046"
```

Ordinal model 1 (all variables)

```
##
## predicted_class_ord1 low risk mid risk high risk
##            low risk        80       42         3
##            mid risk        43       38        28
##            high risk        6       16        49
```

```
## [1] "accuracy=0.547540983606557"
```

```
## [1] "AIC="
```

Ordinal model 2 (blood sugar, systolic blood pressure, age)

```
##
## predicted_class_ord2 low risk mid risk high risk
##            low risk        82       44        16
##            mid risk        45       42        17
##            high risk        2       10        47
```

```
## [1] "accuracy=0.560655737704918"
```

```
## [1] "AIC="
```

Ordinal model 3 (blood sugar, systolic blood pressure)

```
##
## predicted_class_ord3 low risk mid risk high risk
##            low risk        80       48        15
##            mid risk        47       38        18
##            high risk        2       10        47
```

```
## [1] "accuracy=0.540983606557377"
```

```
## [1] "AIC="
```

Ordinal model 4 (blood sugar)

```
##
## predicted_class_ord4 low risk mid risk high risk
##            low risk       115       71        16
##            mid risk        12       15        13
##            high risk        2       10        51
```

```
## [1] "accuracy=0.59344262295082"
```

```
## [1] "AIC="
```

Predicting Medium Risk

	x
Sensitivity	0.9808612
Specificity	0.9166667
Pos Pred Value	0.9624413
Neg Pred Value	0.9565217
Precision	0.9624413
Recall	0.9808612
F1	0.9715640
Prevalence	0.6852459
Detection Rate	0.6721311
Detection Prevalence	0.6983607
Balanced Accuracy	0.9487640

GLM model construction:

we first took the approach of trying to predict the level of risk involved with pregnancy using multinomial and ordinal regression modeling. Starting with all variables and then paring down based on correlation to risk level as well as co-linearity in predictors. we selected first age, systolic blood pressure, and blood sugar level as our first set of restricted predicors followed by blood suger and systolic blood pressure for our second. Although we expected to improve the prediction rate by only keeping the most correlated predictors both model 2(systolicBP, Age, blood sugar) and model 3(systolicBP, blood sugar) in both multinomial and ordinal regressions showed a decrease in predictive accuracy when compared to models with all predictors. The only model in wich we saw the best improvment of our predictions was when we only included blood sugar as a predictor. Because assessing pregnancy risk is very important topic and could leaad to life saving interventions we want to strengthen this model and suggest more in depth analysis into this data via unsupervised learning methods.

Discussion and Conclusions:

Model Interpretation:

First we take a look into multinomial model assessing the accuracy of all predictors and pairing things down to try and improve the model. Starting with the every variable the multinomial model gives a miss error

of 41% which is not great. from here we eliminated redundant as well as with low correlation values to the class level. This did not work out as planned. The first paired down model consisting of Age, Systolic blood pressure, and Blood sugar showed an even worse error rate around 44%. pared down more with blood sugar and systolic blood pressure the missclassification was around 48% and the only improvement was upon making a model solely with blood sugar when the error rate dropped to 40%. I believe that this shows the multinomial regression clearly does not reflect or predict the data too well. One potential future for this type of model would be a boosted multinomial model to help improve accuracy of predictions.

References:

Be sure to cite all references used in the report (APA format).

Appendices:

Supplemental tables and/or figures.

R statistical programming code.