Influential Factors on Women's Education in Indonesia, 1987

Ananya Krishnan, Sara Reynolds, Morgan Nañez
(Dated: May 2020)

This paper explores the factors that best determine what educational level a woman has achieved based on the 1987 National Indonesia Contraceptive Prevalence Survey. We compare and contrast the prediction success or failure for 3 common prediction models: Multi-Class Logistic Regression, Decision Tree, and Random Forest modeling. The features collected in the dataset, originally collected to predict women's contraceptive use, include the age of the wife, their religious affiliation, their contraceptive use, media exposure, standard of living, and the husband's education level. Our development of the final model consists of multiple stages including background research on the study population, initial data exploration, and adjustment for overfitting in our prediction models by tuning hyperparameters. We find that the multi-class Logistic Regression and Random Forest models are the most successful of the three initial models. After analyzing precision and recall for those 2 models, we ultimately choose the multi-class logistic regression model that gives us an accuracy of around 55% in predicting a women's education level in Indonesia in 1987. We conclude by discussing the pitfalls of the data set and the model creation that might have introduced bias or inaccuracies in our prediction.

I. Introduction

Across the globe, women's education has not been a priority; approximately 132 million girls [1] are out of school. Looking at contraceptive data from Indonesia in 1987, we naturally became interested in the education feature and proposed the question, "Within the dataset, what are the factors that best determine a woman's level of education in Indonesia in 1987?" Research on the Indonesian community highlighted some interesting components that informed our assumptions about the data and initial hypothesis about the features that would be most influential on women's education. Most notably, we thought that looking at religious trends related to women's education would be especially important with this dataset because Indonesia has the largest Muslim population at about 87.2% [2]. Access to education for women in predominently Muslim communities is a well known contentious topic, brought to light by Malala Yousafzai and other women's rights activists. We assumed that in extremely religious and typically patriarchal societies that there is also not much emphasis on women's education and that it may be largely influenced by their husband's status (educational level and occupation). Finally, we also predicted that a woman's standard of living, age, and number of children may also be strongly indicative of women's educational level.

To contextualize the data used, the dataset is part of a larger survey from the 1987 National Indonesia Contraceptive Prevalence Survey, which was used to predict contraceptive use. The samples taken were all married women who were either not pregnant or did not know at the time of the survey. Most of the data is categorical and there are 10 total attributes of the dataset: wife's age, wife's education, husband's education, number of children, wife's religion, wife's work, husband's occupation, standard of living, media exposure, and contraceptive use. In this paper, we will propose and evaluate 3 different prediction models: Multi-class Logistic Regression, Decision Trees and Random Forest to evaluate the weight of the aforementioned features in predicting a women's level of education. We will show that the multi-class logistic regression model is the best of the 3 models explored in predicting women's education level in Indonesia in 1987.

[1] `https://www.unicef.org/education/girls-education`

[2] https://www.pewresearch.org/fact-tank/2017/01/31/worlds-muslim-population-more-widespread-than-you-might-think/

II. Data Cleaning

The data cleaning portion of our model consisted of 3 parts. To get a better understanding of the data we were working with, the first was to identify the different data types in our data set. We identified the data types as such:

- Wife's Age→Quantitative Discrete
- Wife's Education→Qualitative Ordinal (Categorical)
- Husband Occupation→Qualitative Nominal (Categorical)
- Husband's Education → Qualitative Ordinal (Categorical)
- Number of Children Born → Quantitative Discrete
- Wife's Religion → Qualitative Nominal (Categorical-Binary)
- Wife's Working → Qualitative Nominal (Categorical-Binary)
- Standard of living → Qualitative Ordinal (Categorical)
- Media Exposure → Qualitative Nominal (Categorical-Binary)
- Contraceptive Method → Qualitative Nominal (Categorical)

As in traditional data cleaning methods, we started our data cleaning by checking for NA or Null values and inputted default values. We checked all columns and found no NANull values. In fact, we observed that all the data was in integer format. Most of the data were categorical(ex. religious) with integers representing a category in a feature and others were numerical (ex. `wife_age`) with integer values over a range. Our discovery of no Null or NA values was further confirmed by the creators of the data set that reported no missing values.

This led to high suspicion that the data might have several default values. However, we found that any suspiciously high count values were consistent with research on the culture in Indonesia. Of the 1473 surveyed women in the data set, 1104 of the women were not working (indicated by a 1 in `wife_work` column) while only 369 were working (indicated with 0 in the dataset). This was also somewhat consistent with our external research that showed that around 50 % [3] of the women in Indonesia do not work. While the percentage of women in the workforce in the data is less than what we found in our research, it is possible that this difference comes from a sampling bias, given the limited size of this sample. We further observe a similar trend in religion, in which the count of those who are followers of Islam (indicated as a 1 in the data set) far exceeded those who were not (indicated as a 0 in the data set). This is also consistent with our research that showed that Indonesia is a predominantly Islamic country.

Our examination of the `value_counts` of each of the columns of the `data_set` also revealed significant outliers and class imbalance that will become essential in our interpretation of the data and model. First, there were few mothers who had more than 12 children. Due to the higher concentration of number of children between 2 - 6, any data point above 12 appeared to be an outlier. Secondly, we were seeing a larger range of wife ages (as this data was not categorical) from 15-50 with the highest concentration of ages between 25-28. Lastly, we observed a significant disparity in the number of women surveyed in each of the categories of `wife_education` and `husband_education`, where 1 represented lower education and 4 represented higher education. There appears to be many more data points in the categories of 3 and 4 in both education of wife and husband as opposed to lower education levels of 1 and 2, which might have played a strong role in our model.

A. Not sure what to name this

The second stage of our data cleaning process occured after our investigation of the data visualized. Due to the qualitative multi-class nominal nature of the `husband_occupation` feature that was inconsistent with other features, it was incredibly difficult to understand the relationship between the categories of the `husband_occupation` column and `wife_education`. For this reason, we decided to remove this column from our analysis. We also decided to change the contraceptive column from ternary to binary categories such that a 0 indicated no contraceptive use and 1 indicated contraceptive use of any kind (long term or short term). As will be explored further in the data visualization section, this change to binary categories amplified the relationship between contraceptives and wife education, namely as wife education increased, so did the proportion of contraceptive use and as wife education dropped, so did the proportion of contraceptive use. We thought this amplified trend from the binary format would prove to be a stronger feature in predicting `women_education`.

B. Data Transformation

The final stage in the data cleaning process was standardization and one hot encoding. Since a majority of our data was categorical, we had one-hot encode our categorical variables to ensure we did not interpret the magnitude of each numerical category differently. The two numerical categories, `wife_age` and `num_child`, were left as is, as the magnitude of their values were inherently important to the data. However, these two numerical features also had to be standardized to ensure `wife_age` and `num_children` values were evaluated on a similar range of -1 to 1. This would also ensure our loss model (defaulted with L2) was proportionally reflective of the actual loss of our model.

III. Data Visualization

We started our data visualization process by trying to understand the relationship between the existing features. Figure 1b depicts the correlation heat map between the features in the data set. The motivation behind creating this plot was to visualize the features with the highest correlation features with one another. We wanted to specifically focus on the highest correlating features with `wife_education` to guide what features might be most interesting to further explore. The highest correlating features with wife education are standard of living and husband education. We explore all of the features with wife education in the following graphs.

A. Proportion Plots

The following 6 graphs are all proportional count plots. Because there is unequal spread of data across each of the categories for `wife_education`, we use a proportional count plot to be able to better compare the data between the different categories.

Figure 1c shows the relationship between Wife Education and the Average Husband Education. As seen in the heat map, the line plot clearly indicates a positive linear correlation between wife education and husband education: as the wife's education level rises, so does the husband's education level.

Figure 1d shows the proportional count plot between `husband_education` and `wife_education`. As the `wife_education` level increases, the proportion of husbands that have a level 4 education increases, while the proportion of husbands that have level 1 or 2 education decreases.

In figure 1e, we look at the distribution of each contraceptive type (1, 2, and 3) per category of wife education (1, 2, 3, and 4). We use the proportion of counts of contraceptive type per category of wife education because we found that the data heavily consists of wife education levels 3 and 4. Instead of comparing the actual count, we compare the percentages of each contraceptive use per education level. While contraceptive use 1 (no-use) declines with higher wife education levels, contraceptive use 2 (short-term use) rises. We don't see a clear trend with contraceptive type 3, leading us to further exploration

[3] `https://www.monash.edu/business/cdes/research/publications/publications2/Womens-economic-participation-in-Indonesia-June-2017.pdf`

with this contraceptive data, which we will look closer at in graph 5.

In figure 1f, we convert the contraceptive type variable (a categorical variable) into contraceptive use (either use or no use) - making it a binary variable. Doing so clearly shows a relationship between education level and contraceptive use. As education level rises, contraceptive use increases and no contraceptive use declines.

Figure 2a indicates that there is a high proportion of religious wives across the board which is likely due to the fact that this data is taken from Indonesia, which has one of the highest populations of Muslims. The greatest non-religious proportion is for wives with an education level of 4. What is interesting in this graph is that the difference in proportion between religious and nonreligious decreases as we increase education levels, indicating that in this data set that religious affiliation decreases with increased education.

In figure 2b, we observe there is little variation in the working status per wife's education level. Consistent within all education groups is that the majority of the women are not working. This graph further informs us that wife work might not be the best predictive feature as there appears to be no trend in the relationship between the two variables being examined.

Figure 2c indicates that the heaviest proportion of not good media exposure is within wife education level 1. The proportion of good media exposure per wife education tends to slightly increase from education level 1(lower) to education level 4 (higher), which matches our intuition that those with higher education might look for more reliable sources of media.

B. Further Investigation

Figure 2f visualizes the distribution of wife age per number of children, for each of the women's education levels. There is an upward linear trend in each education level, indicating that older women have a higher number of children. We can see this by looking at the mean in each box plot distribution. Especially after about 3 kids, there is a consistent upward trend for the `wife_age` mean in each of the graphs. We also observe that despite being split by the education level, each graph shows a similarly consistent positive relationship and range of wife ages for each number of children. This counters the assumption we had that the number of children women have decreases with higher levels of education. From these graphs, we can also see that there are some outliers such as women who have greater than 12 children.

Finally, as indicated by figure 2e , it is clear that there exists a positive linear correlation with standard of living and wife education. As education level increases, the average standard of living also increases. Equipped with a good understanding of the data at hand and their relationship to the `wife_education`, we are prepared to examine different prediction models.

IV. Methods and Experiments

The multi-class classification nature of our central question, inspired the exploration of three distinct models:

- **Multiclass Logistic Regression** → we chose to use the LogisticRegressionCV package from sklearn as it not only implements cross validation (`k-fold = 5`) on our behalf, but also tunes regularization hyperparameters. Since we are fitting a binary problem for each education level, the `'ovr'` `multi_class` argument, along with a `'lbfgs'` solver, was best fit for the problem at hand.

- **Decision Trees** → This problem, we observed, could also be solved by answering a series of questions. This would be best represented in the decision tree model, with hyperparameter tuning due to the proclivity of decision tree models to overfit.

- **Random Forest** → To combat the issue of overfitting in the Decision Tree model, we tried Random Forest, which can model the series of questions but also use bagging and bootstrapping to alleviate overfitting.

Our very initial models, showed preliminary results on the accuracy of train and test data:

- MultiClass Logistic Regression:
  - Training Accuracy: 0.5845
  - Validation Accuracy: 0.5905
- Decision Trees:
  - Training Accuracy: 0.9565
  - Validation Accuracy: 0.49637
- Random Forest:
  - Training Accuracy: 0.9565
  - Validation Accuracy: 0.536

As expected, we see similar training and validation accuracies for the Logistic Regression model — a direct result of the implicit cross validation and hypertuning of the regularization parameters from the LogisticRegressionCV package. In contrast, we see significant overfitting in the Decision Tree model. This is not unexpected as Decision Tree has shown to overfit models. To our surprise, Random Forest too showed significant overfitting in the initial model. In order to adjust for the overfitting, we decided to tune our hyper parameters.

A. Tuning Hyperparameters

Though the package used for LogisticRegression implicitly hypertunes our regularization parameter C, we wanted to observe the variation in the accuracy between the train and validation set as we change C. C values less than 1 resulted in the highest validation accuracy. As seen in Figure 3a, we explored C values less than 1. From this graph, we see that the train and validation accuracy does not change much after a C value of around 0.2. This low C value corresponds to increased regularization. Ultimately, we decided to use the regularization parameter implicitly hypertuned in our final model to ensure we did not introduce uninformed bias in the model.

Another interesting hyperparameter we observed in the Logistic Regression model was the cross-validation parameter. While 5 k-fold cross validation is most commonly used and suggested, we wanted to observe how the train and validation accuracy changed with increased

k-folds. Figure 3b shows insignificant changes in accuracy (y-axis) as we change the cv cross validation folds from 2 to 20. Thus, we decided to stick with a k fold cross validation level of 5.

In the process of hypertuning the Random Forest Model, we looked at 3 integral hyperparameters: max depth of the tree, number of iterations, and the max number of samples. In order to ensure the changes observed in accuracy were due to one hyperparameter alone, we carried the tuned values of earlier tested hyperparameters into the later ones. Let's analyze max depth first. Figure 3c shows the test and validation accuracy as we vary the max depth from 0 to 60. Clearly, we see overfitting as the training accuracy gets increasingly high and the validation accuracy decreases after a max depth value of around 3 or 4. If we had not specified an argument to the max depth hyper parameter, the model would have kept running till all the leaves were pure leaves, making it unable to accurately predict unseen data and thus perform poorly on the test data. The `max_depth` parameter was the most essential in ensuring our model did not overfit.

Another hyperparameter we examined was `n_estimators`. Figure 3d shows the test and validation accuracy as we vary the number of `n_estimators` from 0 to 150. As seen in the figure, we see fluctuations in the data that appear to be consistent between training and validation sets across all the test values of `n_estimators`. Unlike the `max_depth` hyperparameter, there was no clear indication of overturning as the validation accuracy constantly varied with the training accuracy. Since any `n_estimator` value over 15 yielded similar results, we decided to use the default `n_estimator` value of n = 100. Finally, we examined the hyperparameter of `max_samples`. From Figure 3e, we observe that training accuracy and validation accuracy follow very similar patterns and therefore did not need to be manipulated to prevent overfitting.

After hypertuning the Random Forest Model to understand what parameters might help improve the model, we also wanted to test different combinations of the parameters. To test this we looked at the accuracy of 4 variations of the Random Forest Model with different input parameters by calculating the average accuracy of each model over 100 trials. We noticed again, that not specifying a `max_depth` parameter of around 3 will cause overfitting, and that our most accurate Random Forest Model uses `n_estimators = 100` (the default), `max_depth = 3`, and no `max_samples` parameter.

B. Feature Selection

Out of curiosity, we also wanted to see which features were most important in classifying each model. Using the SelectModel package in Sklearn, we identified the features that were most important in each of the models. Figure 4a depicts the weight of each feature in classifying the different levels of wife education in the Logistic Regression model. The features `wife_age`, `num_child`, `media` (media exposure) and `husband_education` of levels 1, 2, and 4 were the most important in classifying logistic regression. Looking at this graph, we observed that the positive weights indicate a higher probability of the point being classified in that category. Similarly, the negative weights indicate a lower probability of the point classified in that education level. Note that negative weights do not specify which other level it will fall into due to the multiclass nature of the problem. As a result, we thought adding two positive weighted features (`husband_education=3` and `standard_living=2`) would have positive effects on classification, especially distinguishing between levels 2 and 3 of women's education.

We similarly analyzed the most relevant features of the Random Forest Classifier. Figure 4b looks at the most important features in this model. The features with the highest weights include `wife_age`, `num_child`, `husband_education=4`, and `standard_living=4`. What was very interesting here is that only higher categories of husband education and standard of living were considered important. We hypothesize that this is a result of class imbalance, in which we see a disproportionate amount of individuals surveyed who had a higher standard of living (category 4) and husband education (category 4).

C. Reassess Hyperparameters

Incorporating our research on tuning hyperparameters of each model, due to its significance of the `max_depth` feature on overfitting the Random Forest Classifier, we set the `max_depth` hyperparameter to be 3. With these adjusted hyperparameters, and selecting the most relevant features of the two models, we achieved the following results:

- MultiClass Logistic Regression:
  - Training Accuracy: 0.5519
  - Validation Accuracy: 0.5688
- Random Forest:
  - Training Accuracy: 0.55314
  - Validation Accuracy: 0.5326

D. Compare Final Models

To further our analysis of the output, we also compared the precision and recall of each category of `wife_education` for the respective models. Figure 5b shows the comparison of precision and recall metrics for our predictions. Something interesting to note here is that we are seeing greater variation in the precision and recall in categories 1 and 2 of wife education. In addition, besides the wife education level of 1, we are seeing the Logistic Regression Model predicting more precisely than the Random Forest Classifier. Ultimately, given the low sensitivity of the question under consideration, we were not too concerned with false negatives and thus put more weight on precision over recall. There appears to be no consistent difference in recall, which is expected due to the high class imbalance between the different categories of women's education as well as the multi-class nature of the problem.

Figure 5a depicts the final training and validation accuracy of our two models: Multiclass Logistic Regression and Random Forest. Taking into account the different success metrics analyzed above, we concluded that the Logistic Regression is the best model to use in predicting women's education in Indonesia in 1987 because it not only has the highest precision but highest validation accuracy as well.

V. Analysis and Conclusions

Based on higher precision and validation accuracy, we ultimately chose the Logistic Regression model as our final model in prediction. The final stage in our analysis was to try the model on the test data. Figure 6 compares the accuracy of the training, validation, and test accuracy of the Logistic Regression Model. After comparing the models, tuning our hyperparameters and comparing our models, we were able to achieve a test accuracy of 55% on the Logistic Regression Model classifying the different categories of women's education.

A big question that followed was: Why is the accuracy so low? We hypothesize that this could have come about from a couple different reasons.

- **Class imbalance**: Figure 1a shows the clear class imbalance in the different categories of wife's education. As the level of education increases, so do the number of data points. This is likely the reason why we saw more accurate predictions for women with higher categories of education as opposed to levels to the lower categories.
- **Very small data set**: our data set was incredibly small. We had only 1473 data points, which was not sufficient enough to generalize the data and make accurate predictions
- **Multi-class Question**: In a binary prediction question, if a data point is far enough from one category, the model knows to classify it as the other class. However, in the multi-class nature of our problem, if it was far enough from one category, there were still 3 other classes that the data point could be classified as, which adds more room for error.
- **Relative Probabilitie**s: Multi-class Logistic Regression measures the probability of each data point being in a particular class. However, if all the probabilities are low, which is what we predict happened in our model, the model will choose the best out of the worst probabilities, further adding to the error that contributed to our low accuracy.
- **Not Representative of Survey**: The data we received was not comprehensive or representative of the data collected during the original survey. For example, in the original report [4], we see that the survey had wife education levels actually go up to a level 5 [] , but this data was not included in the sample provided in our data set.
- **Contraceptive Focused Dataset**: The data collected was intended to predict contraceptive use. While women education might be a factor in predicting contraceptive use, it might not be the same that the features used to predict contraceptive use would be most effective in predicting women's education. This could be a reason we are seeing such a low accuracy as the data is fundamentally intended for a different purpose.

In order to adjust for this low accuracy, we looked at the larger dataset from which our data was derived, the 1987 National Indonesia Contraceptive Prevalence Survey. However the data that was available to us was abstracted by researchers (as explained in bullet point 5 above). Without adequate documentation about how and why they chose particular categories and features over others, it was difficult to ensure consistency.

Another relevant question that follows is: Would the accuracy be acceptable in the context of the question? From the official report, we found that there was a 5th category for higher levels of education (college and beyond). However, only 4 levels of education were described in the available data. Assuming that the abstraction the researchers made was just removing the 5th category, we hypothesize that the dataset distinguished between shorter periods of education (ie looking at differences as small as middle school vs. high school). In other cultures we assumed that predicting education levels would not need to be this granular (ie we could categorize middle school and high school education together). While this may be possible for predicting education levels in other countries and cultures, it doesn't make as much sense in a region like Indonesia where only 16% of people move onto higher education levels [5], meaning we do need to be more granular with how we classify levels of education. Thus, we think that our model should be more accurate with regards to the specificity of our posed question.

However, in the context of the contraceptive data collection, it makes sense that we'd be getting a lower accuracy since we are predicting women education levels from a dataset meant to be used to predict contraceptive use.

A. Relevant Questions

What were two or three of the most interesting features you came across for your particular question?

One interesting feature we came across was `husbands_education`. It stood out because of how highly correlated it was with `wife_education`, confirming some of our early assumptions that a wife's education level might be correlated with husband's education. Secondly, we thought it was interesting that contraceptive use increases as education level increases. However, this feature didn't prove to be very effective in the model. Lastly, the `wife_work` feature stood out to us because it was roughly equal across all education categories (around 20% to 30%), even though we had predicted that it would increase with education level. Because of this, it was not a helpful feature for prediction.

Describe one feature you thought would be useful, but turned out to be ineffective

We thought that `wife_education` would be highly correlated with the husband's level of education and occupation. It follows from our assumption that we made

[4] `https://dhsprogram.com/pubs/pdf/FR19/FR19.pdf`

[5] `https://www.oecd.org/education/education-at-a-glance/EAG2019\_CN\_IDN.pdf`

the assumption that those with increased status of occupation would also have a more well-educated family. However, husband_occupation didn't correlate with `wife_education` or even husband_education. Furthermore, it was difficult to interpret because of the qualitative nominal multiclass nature that was inconsistent with the other qualitative categorical variables.

What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?

Our main limitation from the analysis is the dataset size. Because the dataset was so small, it was hard to make accurate predictions. Another assumption that we made was to convert contraceptive to binary to amplify the correlation with education. It may have been true that keeping the ternary classification of contraceptive use was more highly predictive of wife's education.

What ethical dilemmas did you face with this data?

It was fairly easy to make assumptions about what we might see in the data, given our predisposed thoughts on developing countries and the Islamic religion. This poses an ethical dilemma as no one in our team represents Islamic religion or Indonesian culture. Therefore, we might have introduced bias in the questions and hypotheses that may have shown up in the construction of our model. Furthermore, we think there are some ethical concerns in the lack of diversity of training data — again, there were far more occurrences of women with education levels 3 and 4 which leaves out important data from people of lower educational levels. As seen in the example of Google's poor face recognition for people of color, the lack of diversity can strongly influence the outcome and predictive ability of the model.[6]

What additional data, if available, would strengthen your analysis, or allow you test some other hypotheses?

Having access to more data in general would certainly strengthen our analysis. Specifically, more information about each subjects' household region (suburban/urban), parental education history, languages spoken, and finally, income level might serve as better predictors of the women's education level. Also, we noticed that researchers only collected information about households of married women with husbands which places some constraints on the target population of the dataset. Lastly, having more transparency about what the categories of data actually mean would certainly help us understand the data more and potentially help improve how we manipulate our model. This includes getting more granular data that would enable researchers to choose how to categorize or organize data besides already being given categories.

What ethical concerns might you encounter in studying this problem?

One ethical concern is introducing bias because of assumptions we made based on the background knowledge we have. We can address this by talking to a representative of the area to get a better contextual understanding of data. Another ethical concern we had is around understanding the effects of releasing this kind of analysis and information. We could address this by presenting the information with caution and transparency about these ethical concerns. For example, the associations seen in the data should not be construed as causation. Finally, we also thought about subject privacy during collection of data. In developing and religious communities, questions about contraception, for example, might be considered personal. We could address this by asking for consent to ask questions and publishing data anonymously.

B. Surprising Discoveries

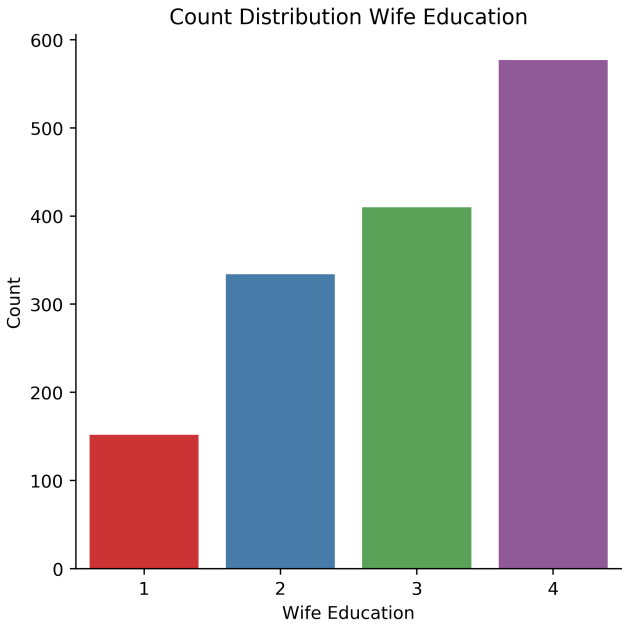
We had a few surprising discoveries as we explored the datasets. First, the positive correlation with `num_child` and `wife_education` went against our initial intuition that women with higher education levels may not have as many children. Secondly, we were surprised that the models placed heavier weight only to some of the feature categories from standard of living and husband education — this is likely due to class imbalance. Finally, we were surprised that the Logistic Regression model considered `wife_work` to be useful even though there was minimal correlation between different categories of wife education or whether or not they worked.

C. Conclusions/Future Work

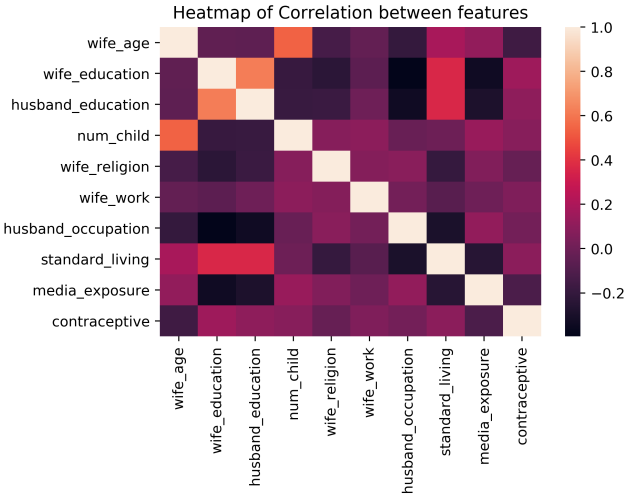
With more data available, we think our models would have performed significantly better, especially if we had equal spread of data across the different categories for women's education. Furthermore, the data we received was not comprehensive or representative of the data collected during the original survey. For example, the report [7] indicates the survey had a wife education level of 5 , which is indicative of higher, collegiate level education, but this data was not included in the sample provided in our data set. The report also indicates that information about other religions besides Islam were accounted for in the survey, but our sample of data only included Islam as a binary feature. In an attempt to add more data, we tried to download the entire 1987 contraceptive data set from the Demographic and Health Surveys (DHS) Program, however we were unable to open .sas7bat files in a comprehensible manner. In the future, it would be interesting to look at other datasets from different countries to compare and contrast features to better understand what factors may help in predicting women's education level.

[6] https://www.cnbc.com/2015/07/01/machine-learning-is-hard-google-photos-has-egregious-facial-recognition-error.html

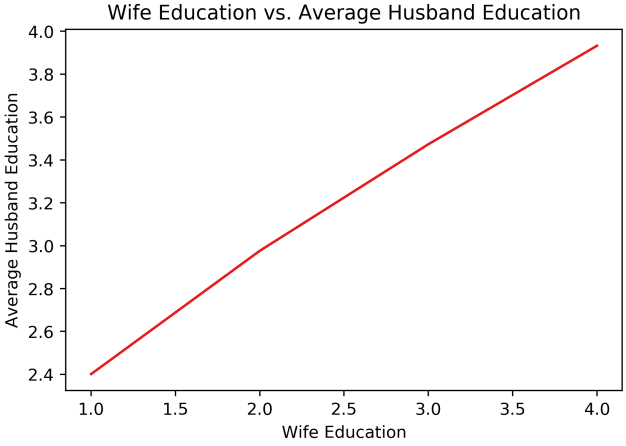
[7] https://dhsprogram.com/pubs/pdf/FR19/FR19.pdf (page 111)



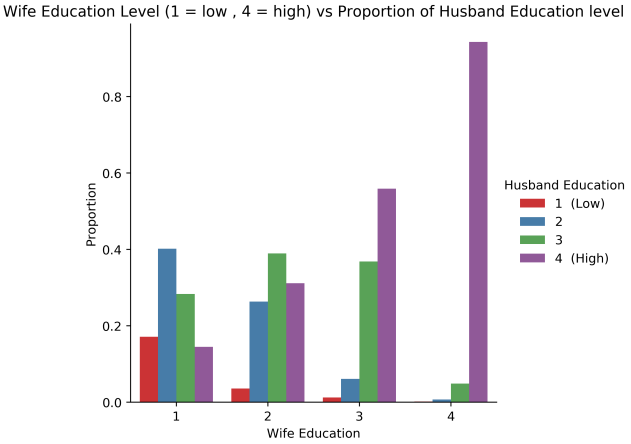
(a) (1) Wife Education Distribution



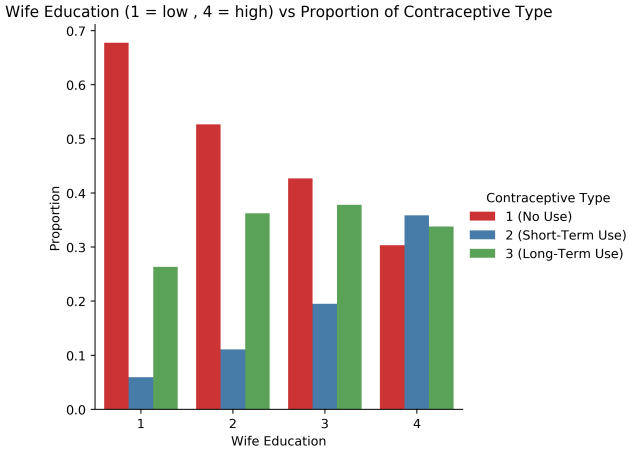
(b) Correlation Heat Map



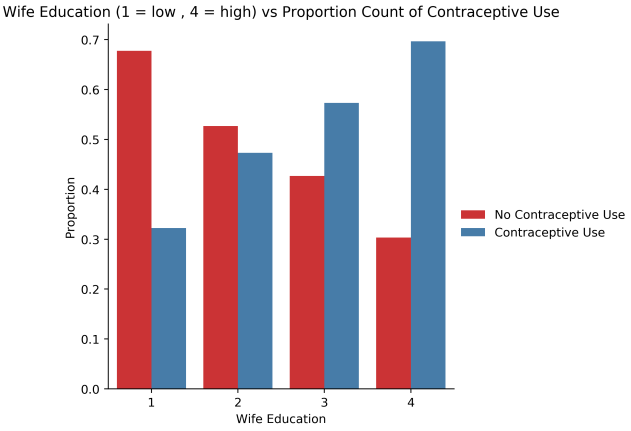
(c) Wife Education vs Average Husband Education



(d) Wife Education vs Husband Education

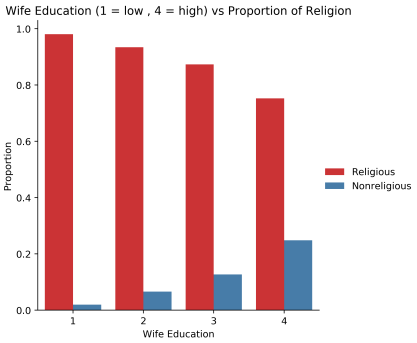


(e) Wife Education vs Contraceptive (trinary)

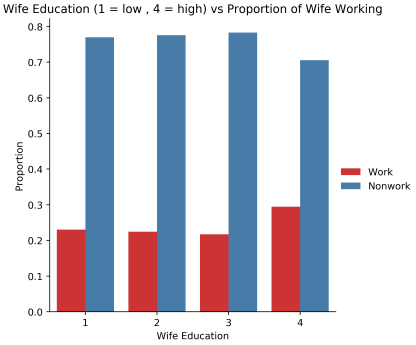


(f) Wife Education vs Contraceptive (binary)

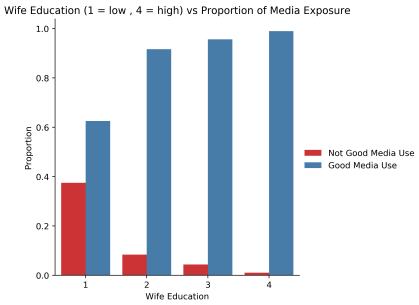
FIG. 1: Exploratory Data Analysis Part 1



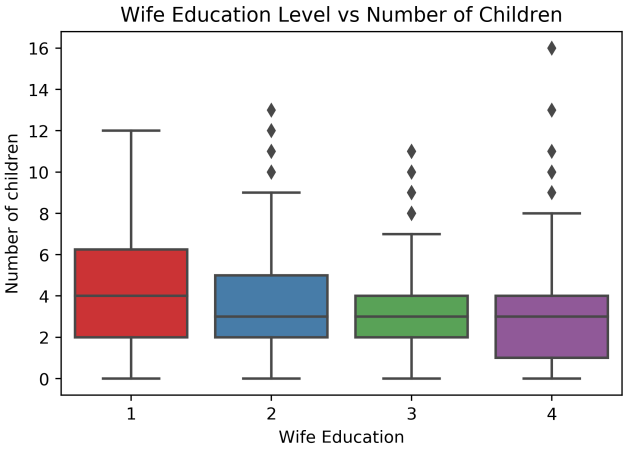
(a) Wife Education vs Religious



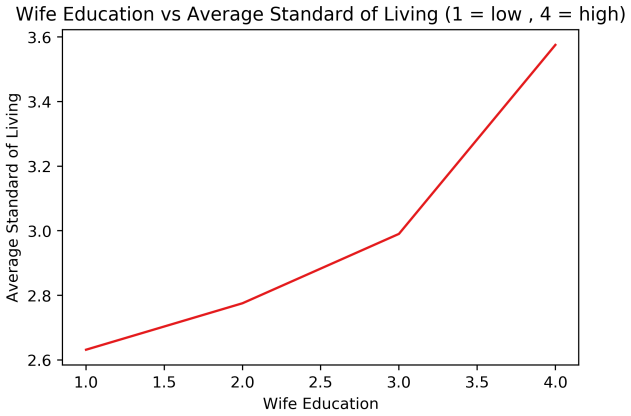
(b) Wife Education vs Work



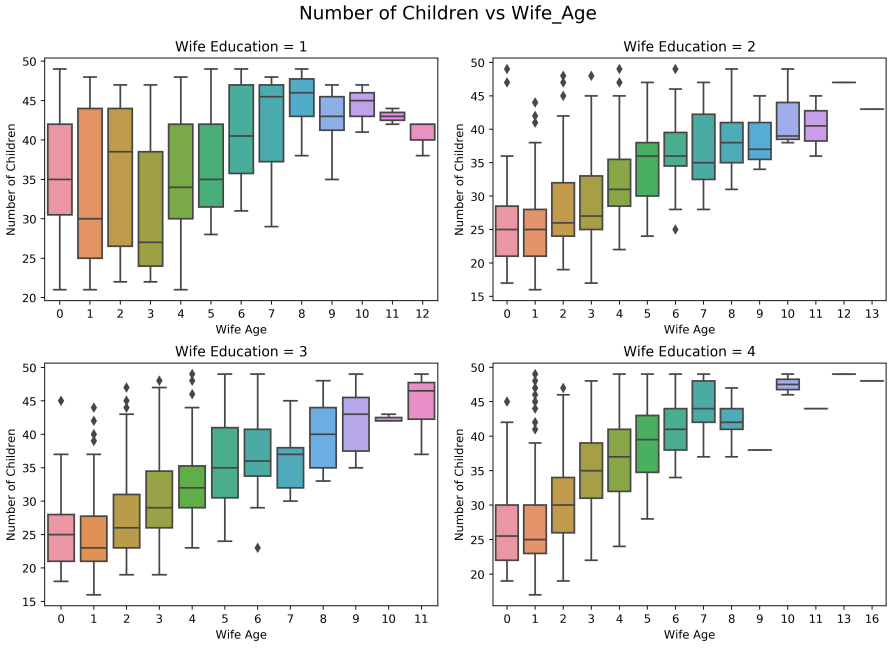
(c) Wife Education vs Media Use



(d) Wife Education vs Number of Children

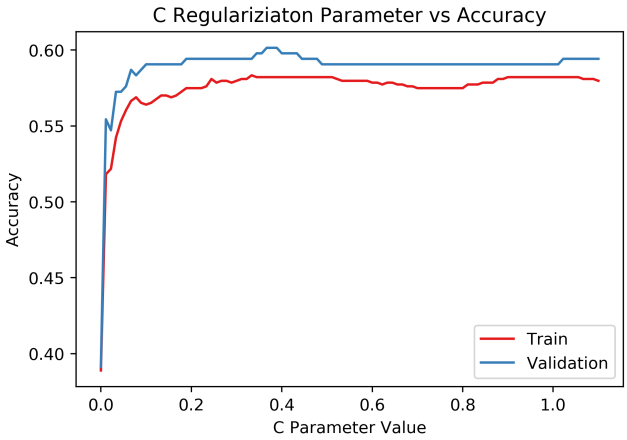


(e) Wife Education vs Standard of Living

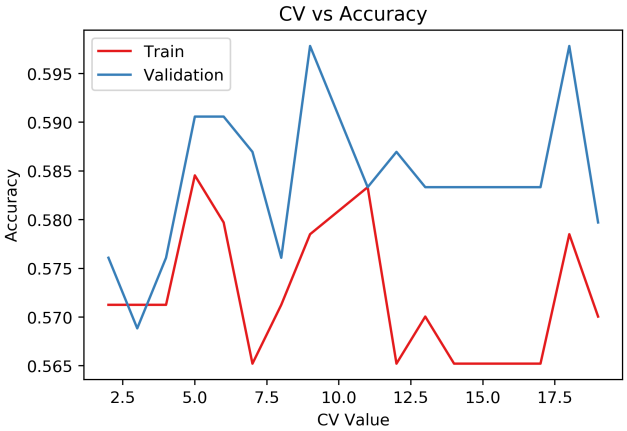


(f) Wife Age vs Number of Children

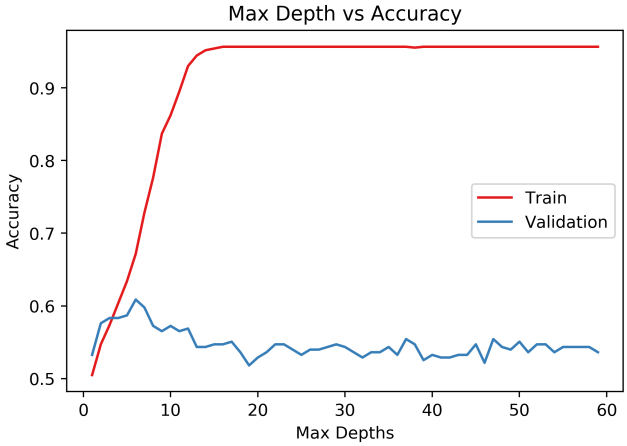
FIG. 2: Exploratory Data Analysis Part 2



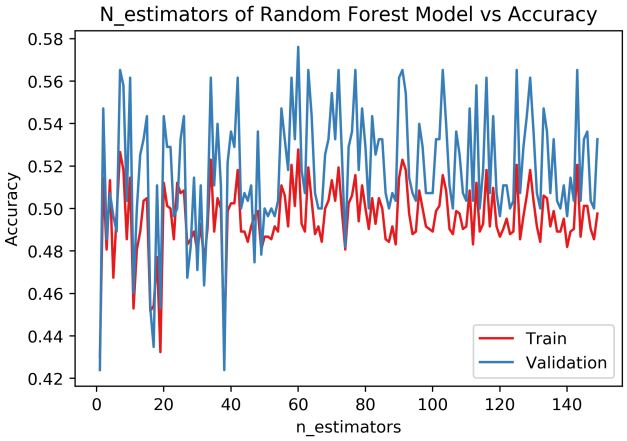
(a) C Regulariziaton Parameter vs Accuracy



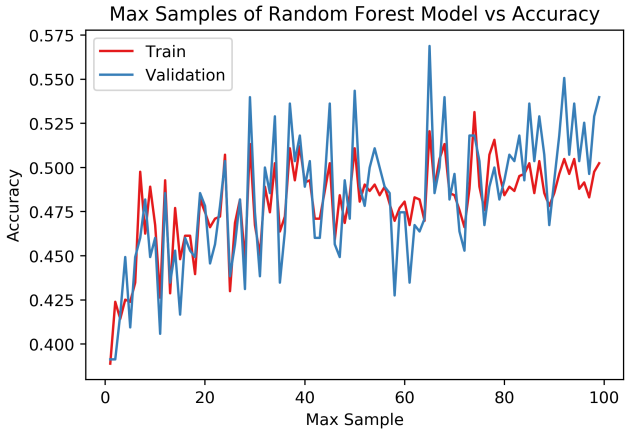
(b) CV Regulariziaton Parameter vs Accuracy



(c) Max Depth Parameter vs Accuracy

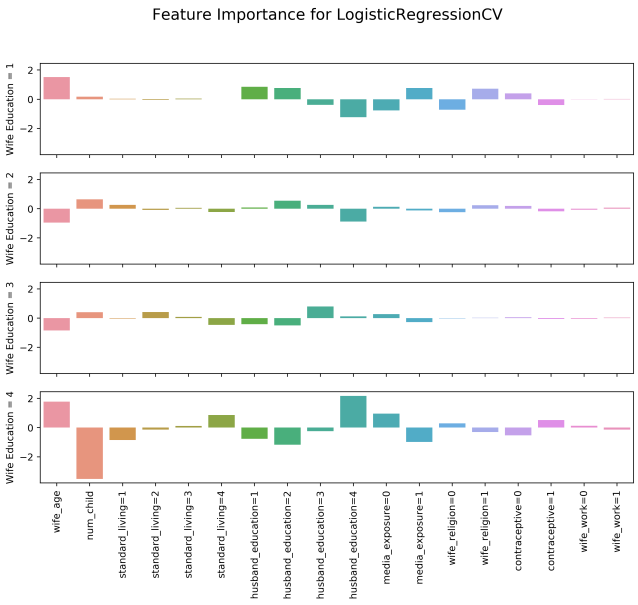


(d) N Estimators Parameter vs Accuracy

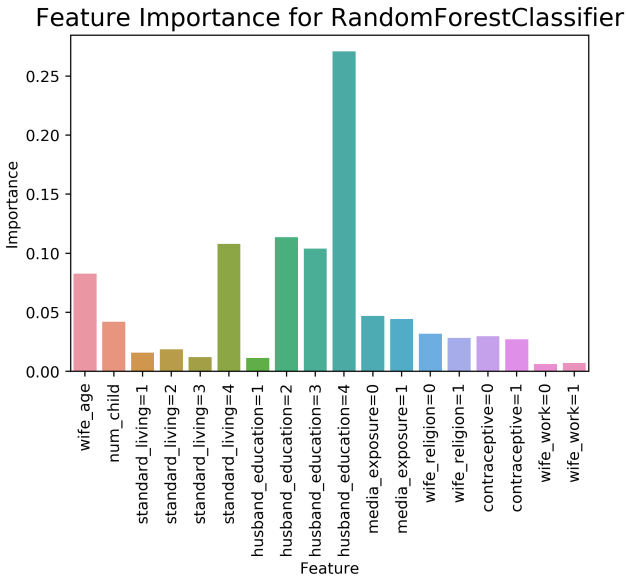


(e) CV Regulariziaton Parameter vs Accuracy

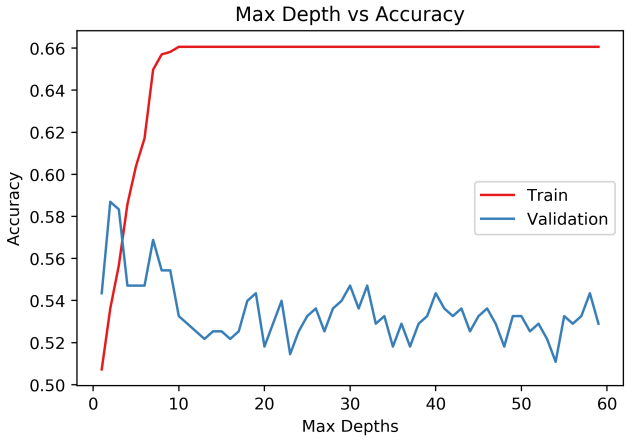
FIG. 3: Exploratory Parameters



(a) LogisticalRegressionCV Feature Selection

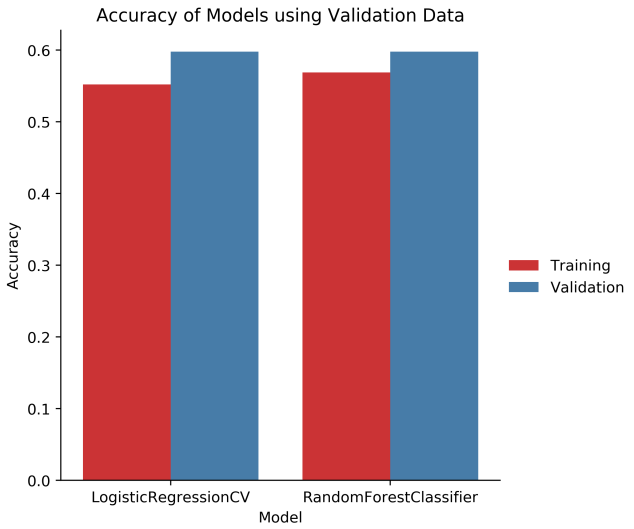


(b) (RandomForestClassifier Feature Selection

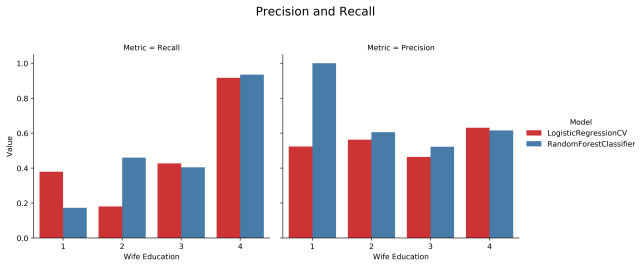


(c) Max Depth Parameter vs Accuracy (reassessed)

FIG. 4: Feature Selection and Reassessment



(a) Accuracy



(b) Precision and Recall

FIG. 5: Compare Final Models Figures 20 - 21

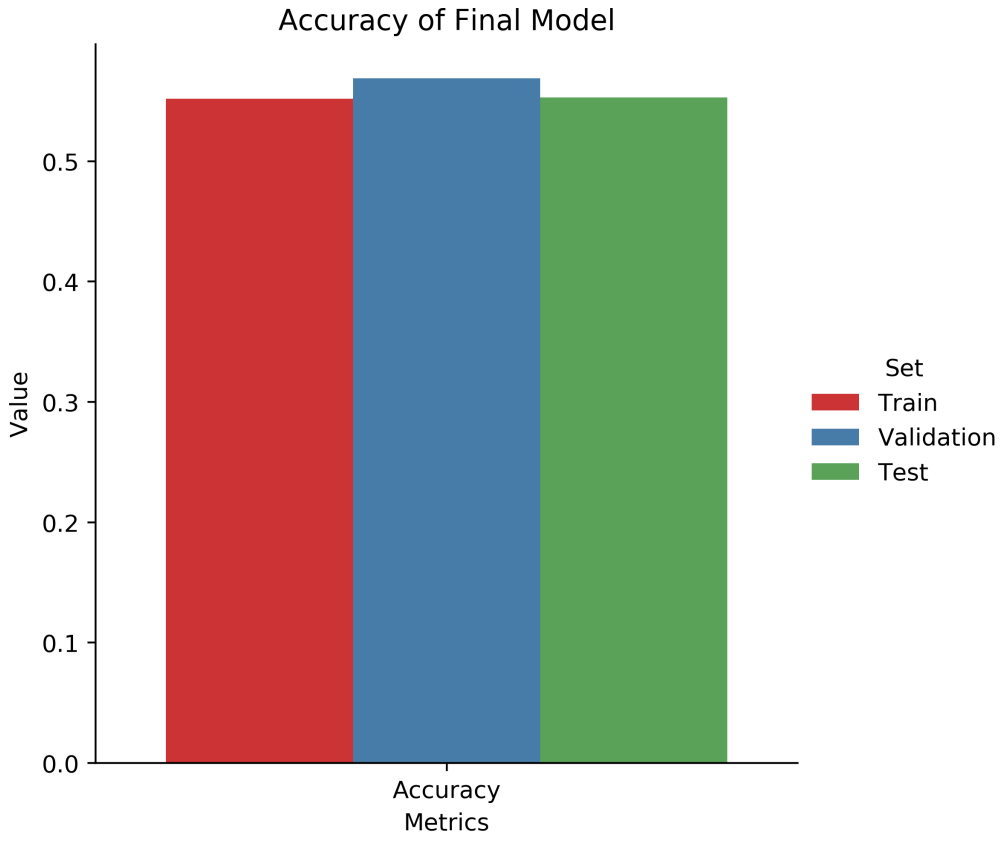


FIG. 6: Accuracy of Final Model