Data Mining for Business Analytics Project:

Targeted Marketing for Women's Health Care

Prakkas Thulasithas
Shashank Naik
Shashank Singhal

1. Business Understanding

1.1 Introduction of the company

Planned Parenthood Federation of America, commonly shortened to Planned Parenthood, is the U.S. affiliate of the International Planned Parenthood Federation (IPPF) and one of its larger members. PPFA is an organization providing reproductive health and maternal and child health services.

Planned Parenthood is one of the largest U.S. provider of reproductive health services, including cancer screening, HIV screening and counseling, contraception, and abortion. Contraception accounts for 35% of PPFA's total services and abortions account for 3%; PPFA conducts roughly 300,000 abortions each year, among 3 million people served.

It is known as a trusted health care provider that delivers vital reproductive health care, sex education, and information to millions of women, men, and young people worldwide.

It is an informed educator, a passionate advocate, and a global partner helping similar organizations around the world. For nearly 100 years, Planned Parenthood has promoted a commonsense approach to women's health and well-being, based on respect for each individual's right to make informed, independent decisions about health, sex, and family planning.

Nationwide they operate approximately 700 health centers, which reflect the diverse needs of their communities. These health centers provide a wide range of safe, reliable health care — and the majority is preventive, primary care, which helps prevent unintended pregnancies through

contraception, reduce the spread of sexually transmitted infections through testing and treatment, and screen for cervical and other cancers.

1.2 The Business Case

The aim of this project is to preempt the requirements for personalized birth control and safe sex counselling for people to increase profits by getting more customers to enroll for their services. As stated above, contraception accounts for a huge chunk of the services. Also, it is an entry point for most customers using other services provided by them. Other than that, this helps them take a step further towards their motive as a company, which is to educate about unsafe sex and unplanned pregnancy.

The company plans to use two types of marketing strategies each having different costs and a different probability of customer conversion. The cost is assumed to be $15 for one strategy and $7 for the other, and the probability of conversion of a person to a customer varies is higher if strategy 1 is used as compared to strategy 2.

We are the data science team for Planned Parenthood. Using data mining we will identify people who are likely to require or opt for birth control or sex related counselling over the next 6 months. Planned Parenthood will then market their counselling services to these people which should result in higher customer enrollment and thus, higher profits. According to which marketing strategy gives highest returns with groups of people having different probability of joining the service, we will try to maximize profit. The main decision to be made is selecting the marketing strategy for the respective subset of people.

2. Data Understanding

We acquired the data from a Data Mining competition on DrivenData.org. The data is based on the NSFG (National Survey of Family Growth) public dataset conducted by the CDC (Center for Disease Control and Prevention). People across the United States were asked a series of over 1700 questions about their demographics, pregnancies, family planning, use of healthcare services, and medical insurance. We're focusing on the respondents to these questions that are women, and each row in the provided data represents an individual. To give you a sense of the kind of data collected, the relevant sections of the NSFG survey are:

Section A - Demographics, household, childhood

Section B - Pregnancy, birth, and adoption

Section C - Marital and relationship history

Section D - Sterilizing operations and impaired fecundity

Section E - Contraceptive history and desire for pregnancy

Section F - Family planning and medical services

Section G - Birth desires and intentions

Section H - Infertility services and reproductive health

Section I - Insurance, residence, religion, work history, child care, attitudes

The data we acquired was obfuscated to maintain its integrity. The feature names indicate the variable type. Numeric variables start with n_, categorical variables start with c_, and ordinal variables start with o_. To access the actual feature names, we contacted the DrivenData team and they helped us out by providing the reversing dictionary which enabled us to better understand what features have predictive power.

Example data instances and attribute names:

Feature data types

Each column is one of the following datatypes, which we indicate with the column headers:

Data type	Header prefix	Description	Values	NaN value
Categorical	`c_`	Variables that can take on a set of values which are not numeric or ordered, e.g. yes/no answers	A *case sensitive* single letter	Empty
Numeric	`n_`	Variables that have a numeric value, e.g. number of children	Scaled to the range [0, 1], or all 1 for singleton values	Empty
Ordinal	`o_`	Variables where the order matters, e.g. level of education	Integer value	Empty

Feature example values

Here are a few rows of a selection of random columns of different data types:

	n_0002	o_0140	c_1370
id			
11193	0.025449	11	NaN
11382	0.031297	11	a
16531	0.024475	NaN	a
1896	0.041694	27	NaN
18262	0.038120	17	b

The data is relatively sparse which makes it even more challenging.

It is provided in 3 Comma separated value files.

Training set values: The predictors, or independent variables. These are the features that we use to predict the probability of using the service. There are **1379 features** and **14644 data instances**. Off these we kept **10644 data instances** for our **in sample training data set** and the rest **4000 were kept as out of sample**.

Training set labels: The dependent variables. These are the labels that we build a model on to predict. The instances are mapped to the labels (service target variables). We selected one of the targets for our purpose (service A).

Test set values: The predictors, or independent variables for the test set. These are the features that we use to make the predictions that you submit. There are **4000 unlabeled data instances**.

As discussed in class, to avoid over fitting we need to make sure we have adequate features, instances and base rate. The data provided satisfies all the three conditions.

3. Data Preparation

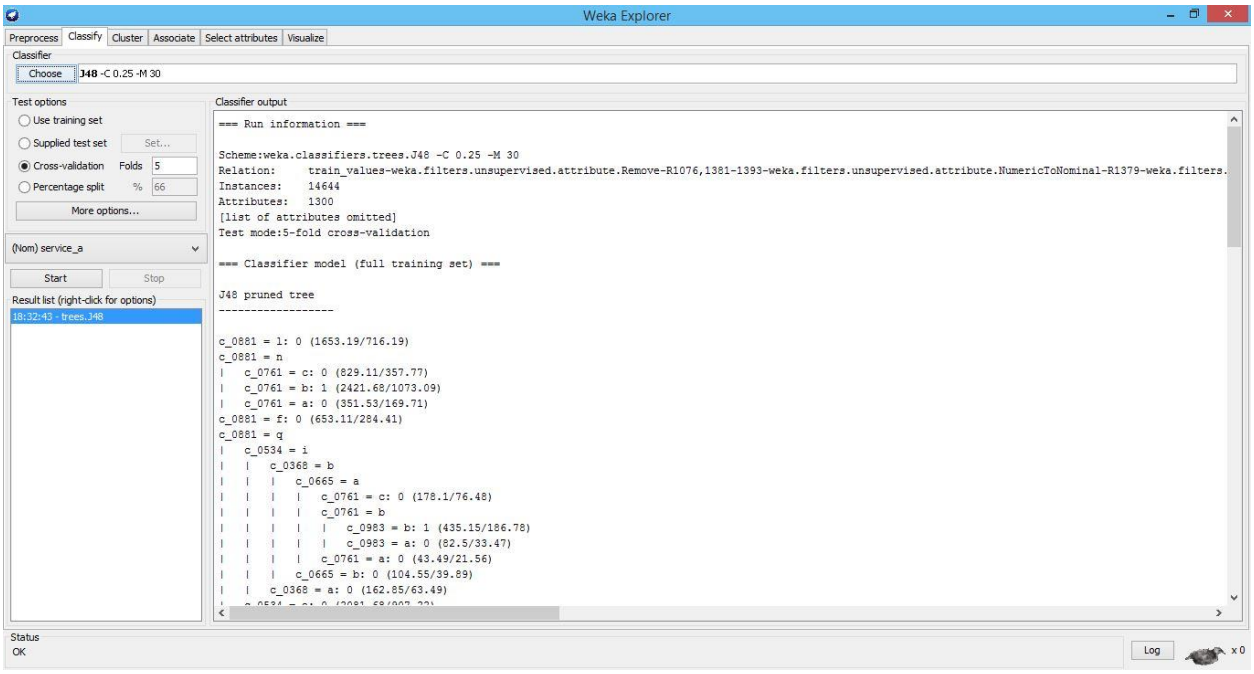
Our target variable being the use or not of service A, we needed to extract that data from the training set labels file to incorporate it in the training set values file. We sorted the data by patient ID and merged the training set values and labels files in Excel.

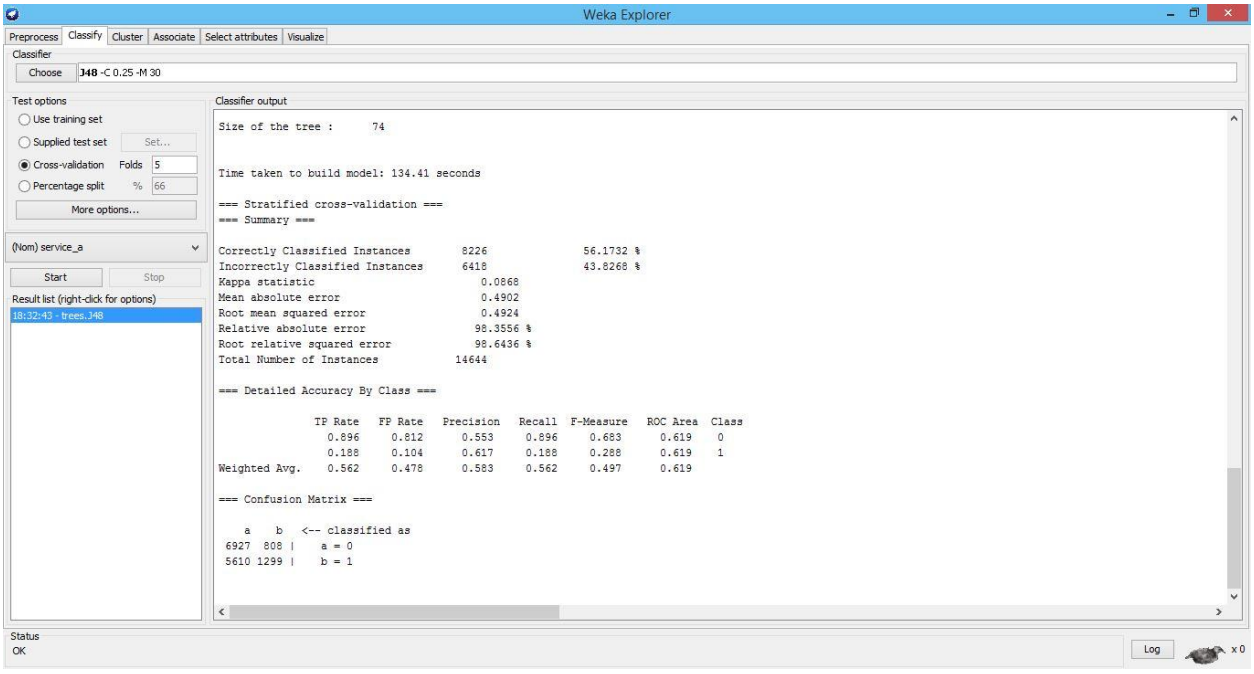
When we loaded our csv files into Weka, all the attributes were automatically recognized as numerical. As we have seen before, our dataset contains categorical and ordinal data then we preprocessed the data using the NumericToNominal filter to correct the assigned data type.

After going through the data we noticed that several features had more than 90% missing values while some other features had the same value for all the examples. We removed the attributes that had same values throughout.

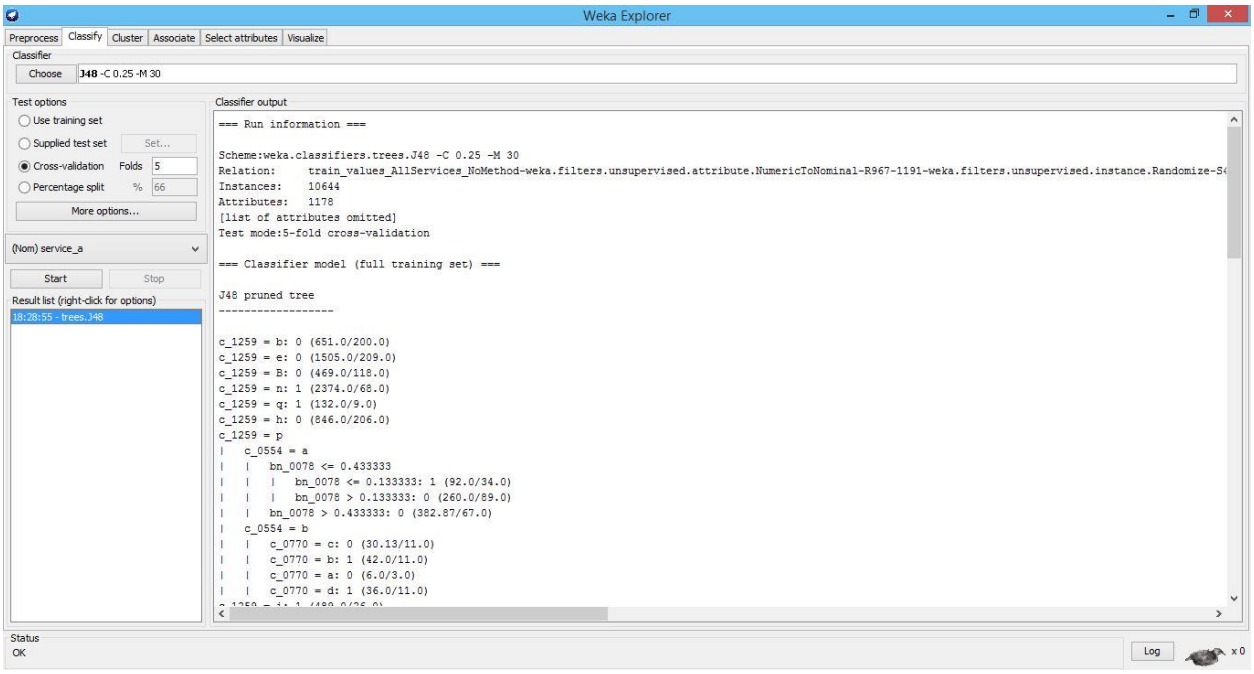
Then, we used the Attribute Evaluator section in WEKA to find the most discriminative features by using the InfoGainAttributeEval function which sorts the attributes according to the information gain. We removed attributes with more than 90% missing values that had the lowest info gain for our target.

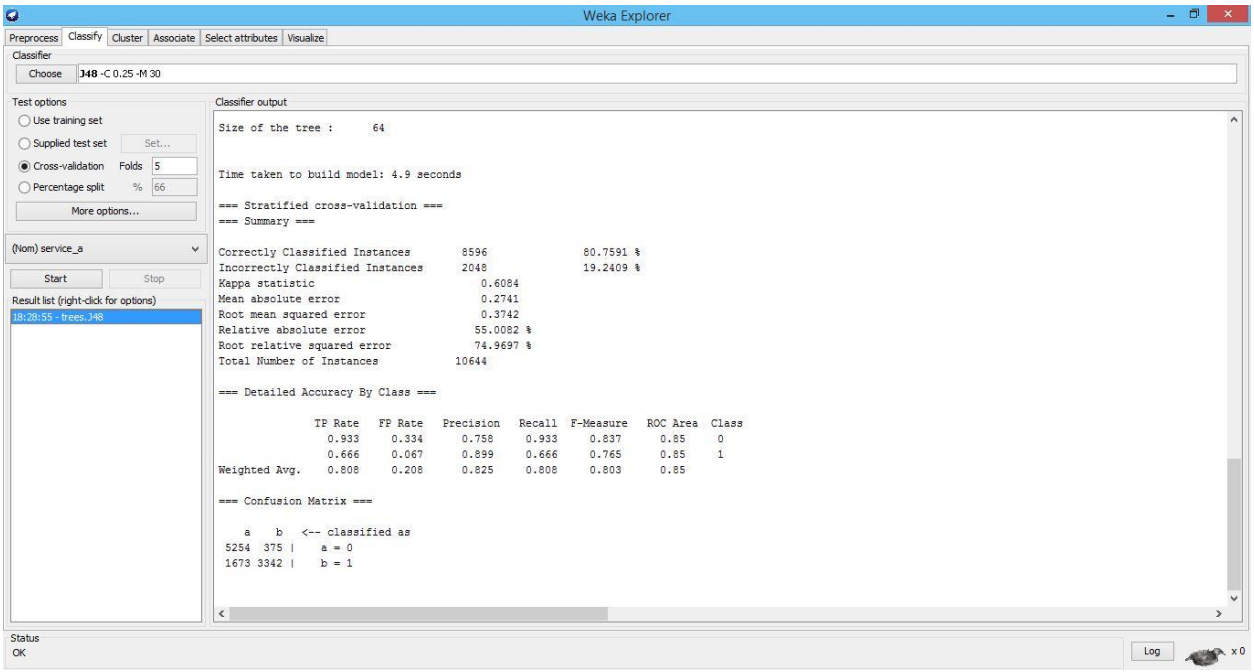
Model built including attributes that have more than 90% missing values for instances:





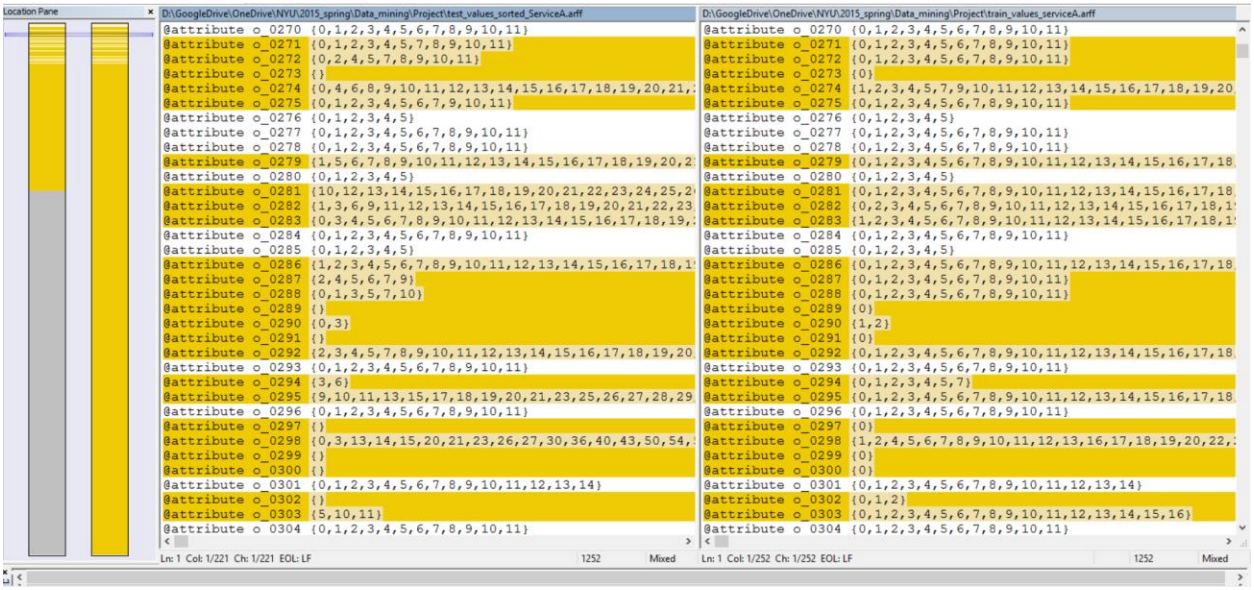
Model built after removing attributes with more than 90% missing values for instances:





Once we finished building our model with our training set, we tried to use the provided test set in order to see how well our model would perform with out-of-sample data. However, when we tried to do so, we got an error message telling us that the train and test sets were not compatible. Obviously, we verified that the structure of both datasets were identical with the same number

of attributes and the exact same type but we still had the error message. In fact, as we can see in the screenshot below by using WinMerge, the attributes in the training and test had different ranges and thus, there were not correctly matched. Therefore, we modified those ranges to be able to supply the test set and evaluate our model.



4. Modeling

Since our target variable is predicting the probability of a patient using a birth control or safe-sex related counseling, we are dealing with a supervised ranking problem so we can implement a targeted marketing campaign.

We built a bunch of models from Decision Trees, Random Forest to Naives Bayes. We couldn't use logistic regression as a significant amount of our features were categorical/ordinal. The models we tested were:

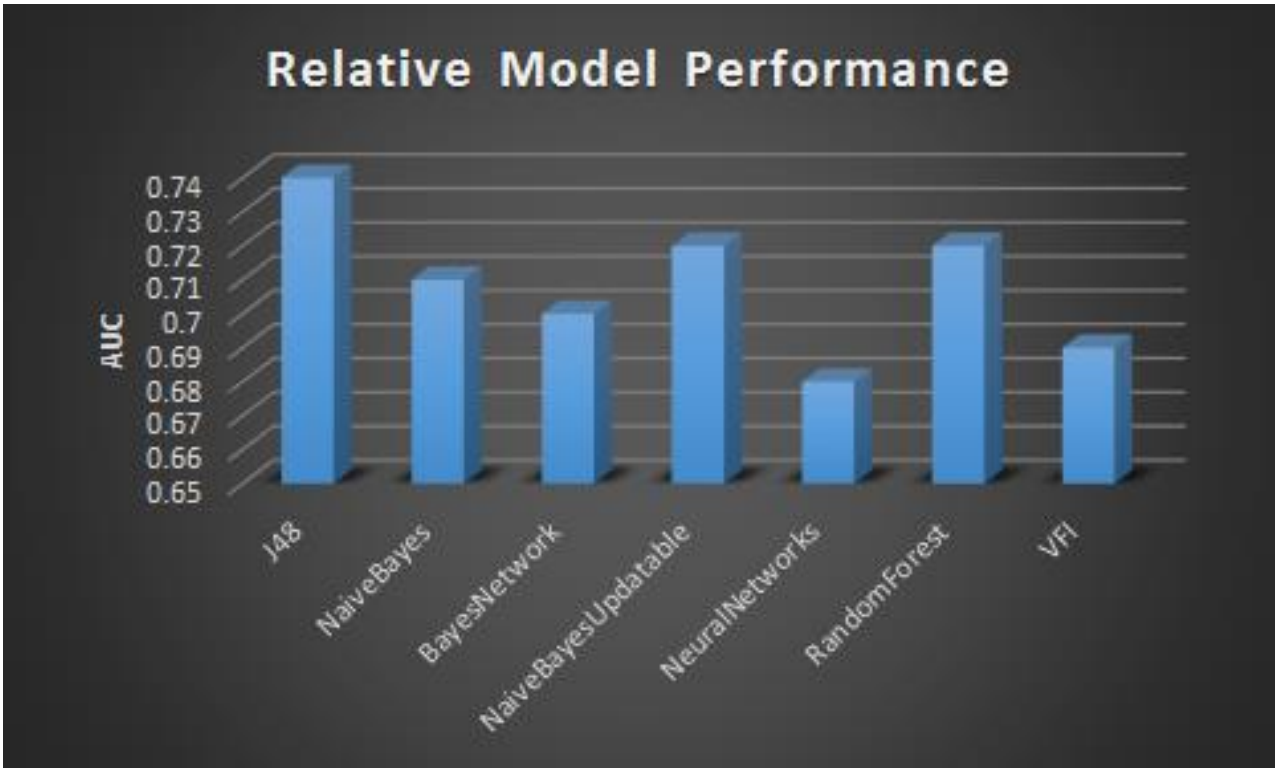
- J48
- NaiveBayes
- BayesNetwork
- NaiveBayesUpdatable
- NeuralNetworks
- RandomForest
- VFI (Voting Feature Intervals)

Also ran the model for SVM (Support Vector Machine), ADT (Alternating Decision Tree), KNN but Weka either took too long or ran out of memory and hence they were not practical for the scope of the project. ADT gave really good AUC, but again running time and memory requirements made it impractical and we guess that the high AUC was due to the high gain attribute boosting nature of the algorithm.

In the chart below, we plotted the AUC we got for the different models we used.

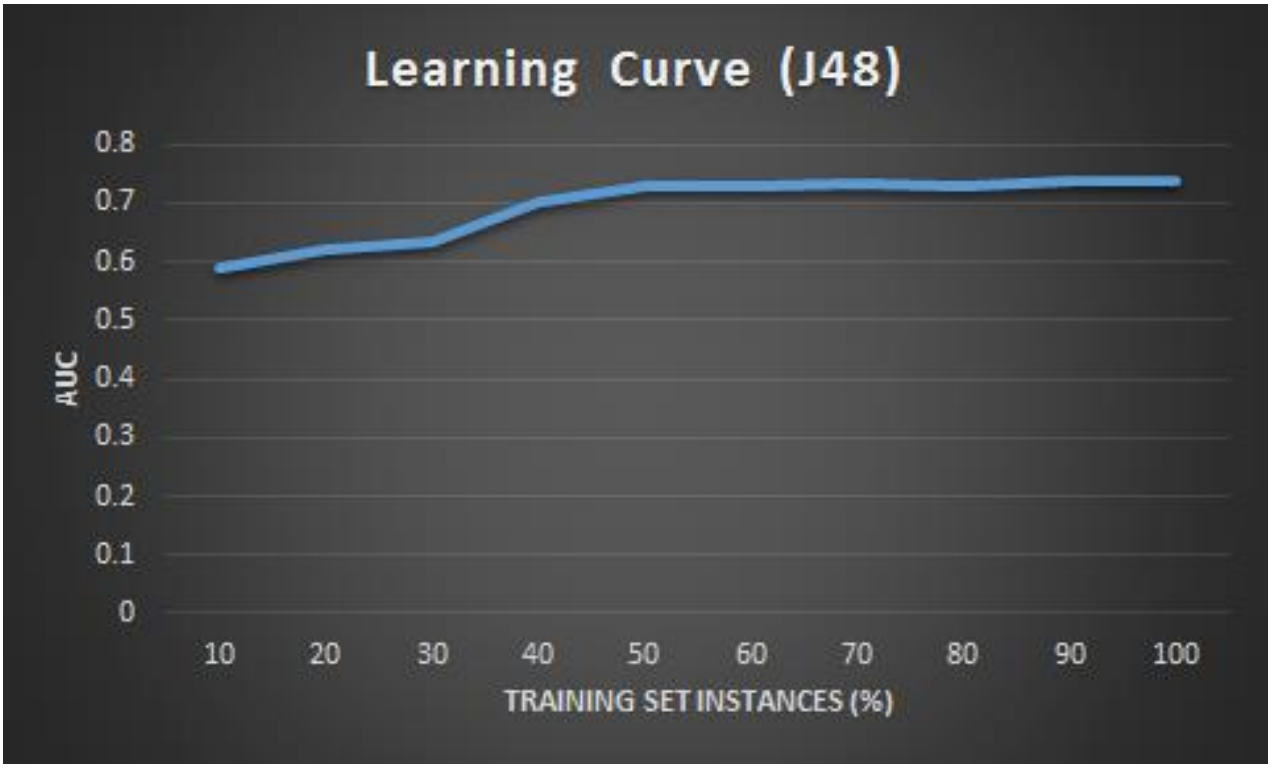
In the end we decided to continue to J48, since it's a simple algorithm and one which we are familiar with.

Comparison between models based on AUC:



Then we ran this model with different values for the minNumObjects parameter and finally, as seen from the fitting curve, we decided minNumObjects = 30 gave us the best performance.

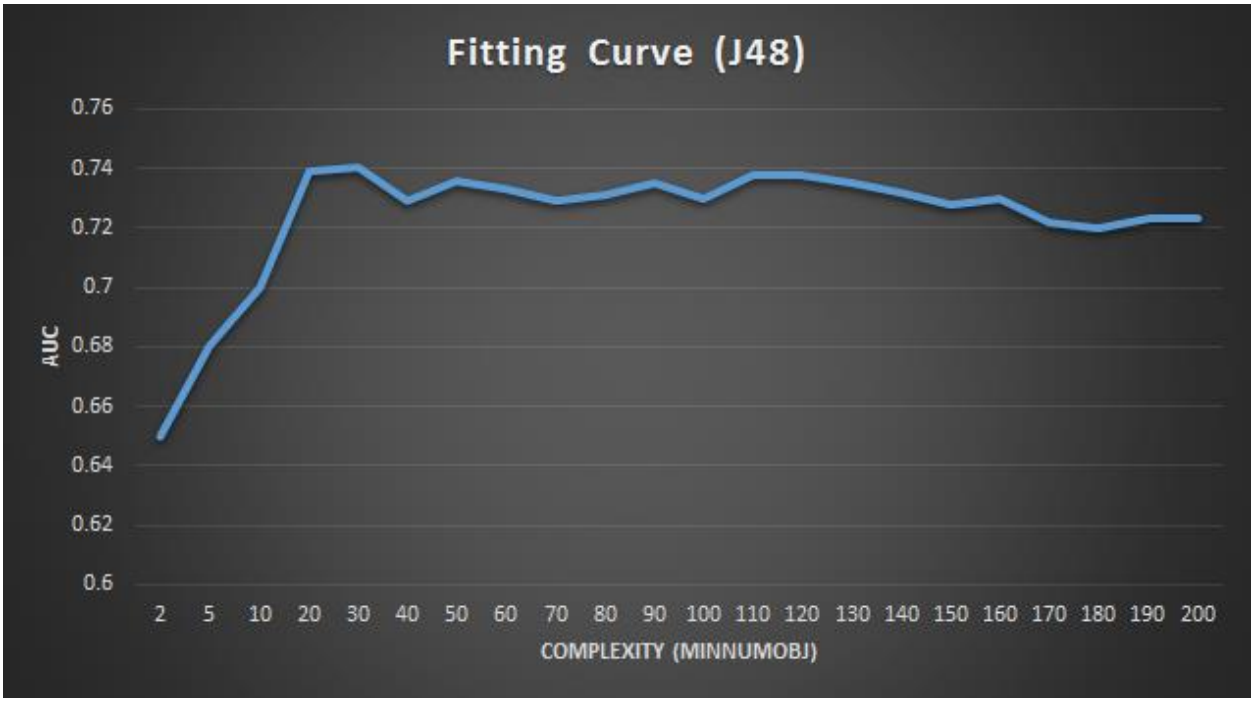
Learning Curve



The above graph is a plot of the out of sample AUC against the size of the training set. As expected, we can see the performance increases as the number of training instances increases. Below is the data we gathered for the Learning curve:

Percentage (Training Set)	**AUC**
10	0.59
20	0.623
30	0.634
40	0.7
50	0.728
60	0.73
70	0.734
80	0.728
90	0.739
100	0.74

Fitting Curve



The above graph is a plot of the out of sample AUC against the complexity of the selected model (J48). As expected, we can see the performance suffers due to over fitting for a low MinNumObj (M) value.

Below is the data we gathered for the Fitting curve:

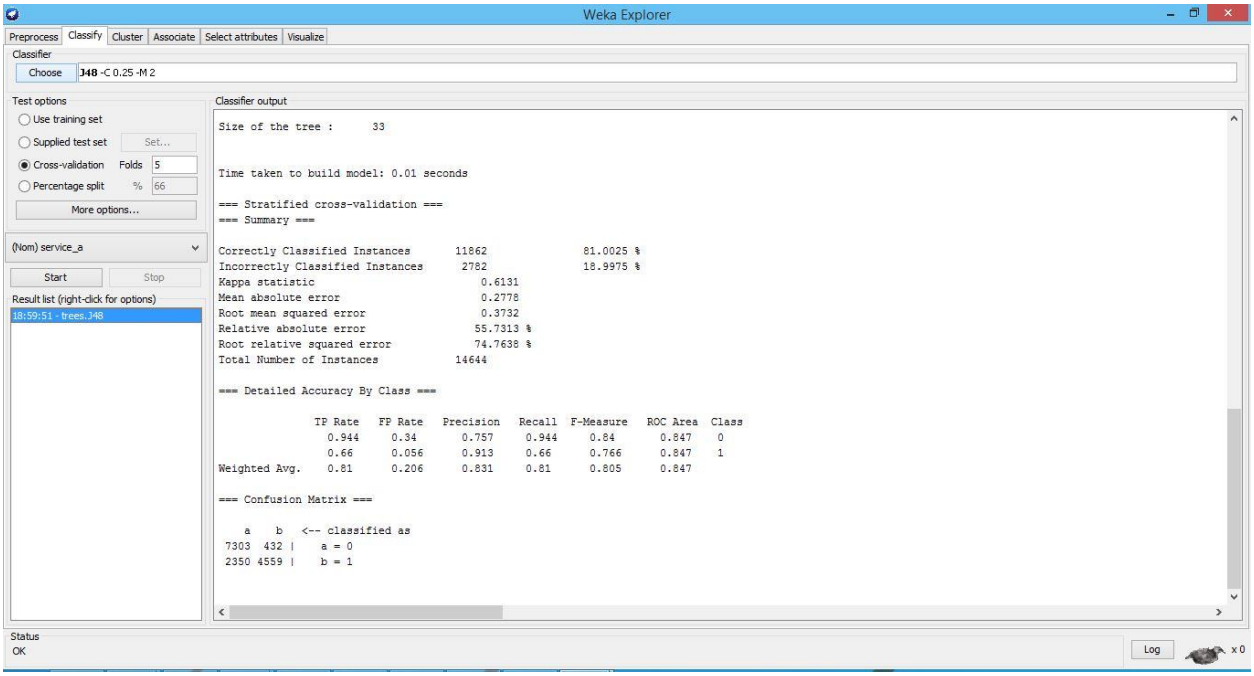
MinNumObj	AUC
2	0.65
5	0.68
10	0.7
20	0.739
30	0.74
40	0.729
50	0.736
60	0.733
70	0.729
80	0.731
90	0.735

100	0.73
110	0.738
120	0.738
130	0.735
140	0.732
150	0.728

160	0.73
170	0.722
180	0.72
190	0.723
200	0.723

Data Leakage

During attribute selection we encountered something strange with the results, there was one attribute (C_1259) with an unusual infoGain that was influencing the target variable almost singlehandedly. We ran the model only with this single attribute and as we hypothesized this attribute on its own was able to achieve an AUC of almost 0.9.



We decided to investigate the issue, we contacted the competition organizer and got the labels of the attributes. We immediately saw the issue smelled like a data leakage. The feature was “Most recent contraceptive method prescribed” and the target is “Counselling for Birth Control or Prescription.” As we can see there is a direct relation between the two. So we decided to remove this attribute to fix the data leakage.

5. Evaluation

The test set provided by the competition didn't include the actual predictions of each instance. Therefore, to evaluate our model, we created our own test set by removing instances from the training set. As a result of doing this, we obtained a 4000-instances test with a baserate of 47% which is similar to the baserate of our training set.

We saw in the previous modeling section that the J48 classifier was the best performer among all the models we tried. Our dataset had a balanced baserate (47%) therefore, we decided to use confusion matrixes to set the correct cut-off and be able to rank the probabilities of whether a person is likely to use our service. As expected, we got the maximum accuracy for a 0.5 cut off whereas using 0.3 or 0.7 gave us, respectively, more false positives or false negatives.

Cut 0.3	Predicted Class	
Actual	1	0
1	1658	236
0	1306	800

Accuracy: 61%

Cut 0.5	Predicted Class	
Actual	1	0
1	1219	675
0	640	1466

Accuracy: 67%

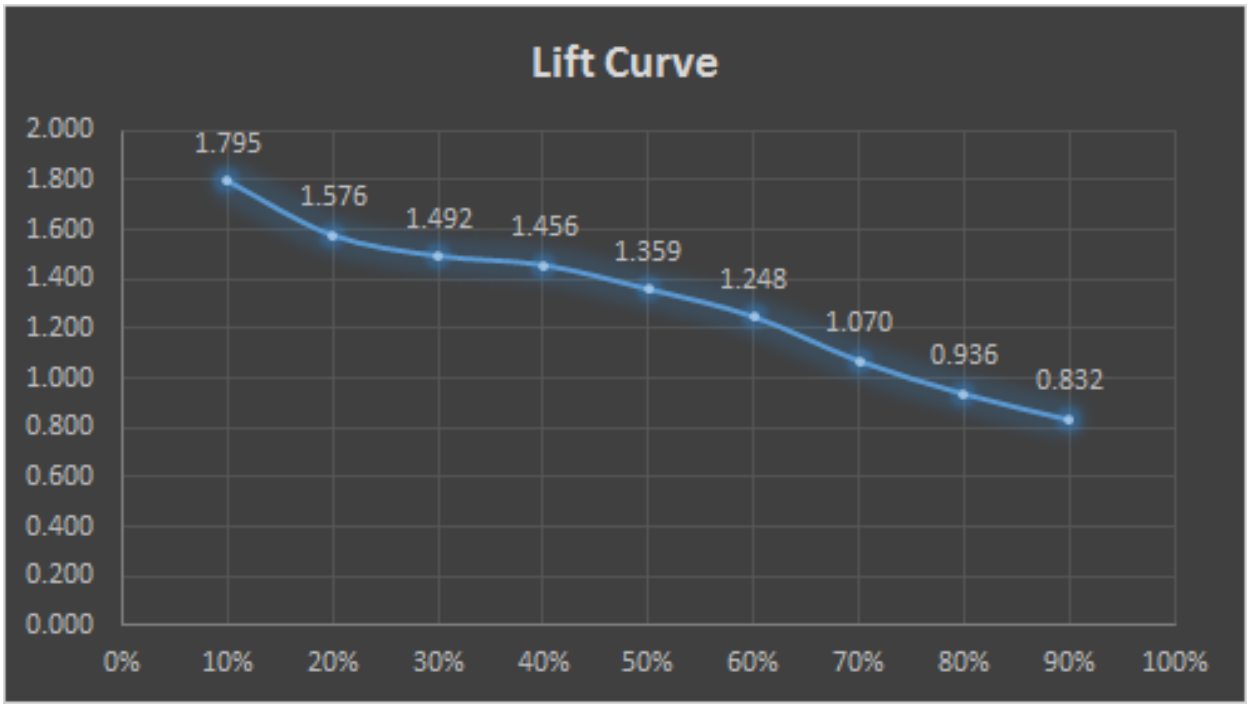
Cut 0.7	Predicted Class	
Actual	1	0
1	283	1611
0	24	2082

Accuracy: 59%

The way we can evaluate business case improvement is as follows:

The model is built on the training data set, based on which we can evaluate the attributes and create a private dataset on those attributes to target the customers for marketing.

In order to present an intuitive visualization of our solution to our client, Planned Parenthood, we decided to plot a lift curve of our model to show how well our model performs in targeting potential customers rather than selecting them randomly. Given that we have a base rate of 47% in our test set, quite half of the customers are likely to use the birth control counselling service and we want our marketing efforts to reach them in priority. As shown in the lift curve below, the less people we target the higher is the lift. Therefore, we will advise to target the top 40% of the patients which represents a lift of 1.45



We have two Marketing offers say M1 and M2, with cost : C1 > C2

Based on the results of the model, we rank the customers on the probabilities of them using the service. Customers above a high cutoff of (0.8) are very likely to using the service and Customers between 0.5-0.8 are moderately likely to be interested in using the service.

We analyze the effects of the two marketing strategies and try to maximize the profit.

P(M1) = The probability of a person converting after being offered Strategy 1

P(M2) = The probability of a person converting after being offered Strategy 2

P(H) = Predicted probability of a person to use the service is more than 0.8

P(L) = Predicted Probability of a person to use the service is between 0.5-0.8

Our goal is to maximize the profit and optimize Marketing expenditure. We can try 4 possible combinations to try optimize the marketing strategy.

Market plan M1 to customers belonging to H group: P(M1/H)

Market plan M2 to customers belonging to H group: P(M2/H)

Market plan M1 to customers belonging to L group: P(M1/L)

Market plan M2 to customers belonging to L group: P(M2/L)

Suppose, M1 is marketed to H and M2 is marketed to L. Then total cost of C1 would be n(H) * C1 and total cost of C2 would be n(L) * C2. The accumulated cost would be

[Total Cost C1+Total Cost C2] = n (H) * C1 + n (L) *C2

Revenue = [P(M1/H) * n(H) + P(M2/L) * n(L)] X (Cost of service A)

Profit = Revenue – Costs

Predicted Results:

Following are the results we got after running the model on the Test set of 4000 instances:

Marketing Strategy	Cost	Probability with H	Probability with L
M1	15	0.5	0.3
M2	7	0.4	0.2

	Number	Cost M1	Cost M2
above threshold of .5	2327	34905	16289
H (above .8)	147	2205	1029
L (.5 to .8)	2180	32700	15260

Assuming the cost of the service is $50, we calculated the highest expected profit was achieved by using marketing strategy 2 for both the subsets. This output is susceptible to change depending on the assumptions made on the price and probability of the strategies.

	M1 H, M1 L	M2 H, M2 L	M1 H, M2 L	M2 H, M1 L
Marketing Cost	34905	16289	17465	33729
Expected number of conversions	727.5	494.8	509.5	712.8
REVENUE	36375	24740	25475	35640
Profit	1470	8451	8010	1911
Profit per person	2.020618557	17.07962813	15.72129539	2.680976431

6. Deployment

The following strategy should be followed for deployment of the solution. The target population will be ranked based on the predicted probability of requiring the service. The population will then be divided into two subsets. Subset 1 will have the people with predicted probability higher than 0.8 and the second subset will have people with probability between 0.5 and 0.8. Both the groups will be subjected to different marketing strategy. Based on the conversion probability given by the marketing team, the marketing strategy that maximizes the expected profit will be used for the target customer. As the conversion probability of both marketing strategies is different and also since it is an estimate, we will keep track of how the marketing strategies are performing. Also, we will be able to find out how good the model predictions are. We will track if a customer opts or shows interest for our service within the next 6 months of the date of the target advertising for this evaluation.

We have assumed that the revenue the company gains from every customer is a one-time pay of the cost of the service. This assumption may or may not be true and the cost can be modified to include the synergy benefits due to customers using other services when they opt in for the target service. Also, it is necessary to monitor the response of customers to different marketing strategies as they are best estimates of what may happen. A change in probability of conversion depending on which subset a person belongs to may cause a shift in choice of marketing strategy to use. The difference in costs of the two marketing strategies also play a big role in this decision as they have a direct impact on the expected profit. Any change in the prices should be updated to evaluate the decision.

Since the Planned Parenthood is a public health care company affiliated to the International Planned Parenthood Federation and has access to link data on public surveys with the actual person, through which it is able to target market to these people, it needs to be cautious how it is using this data. This campaign is about getting more people to have counselling sessions, so it

may not be crossing the moral line, but using the same data to predict and target market about specific diseases or other personal matters may be termed as unethical. Ultimately it is the company's call to decide where the fine line between helping and stepping beyond the comfort zone lies.