CornellEngineering
Operations Research and
Information Engineering

ORIE 4741

Learning With Big Messy Data

Final Report

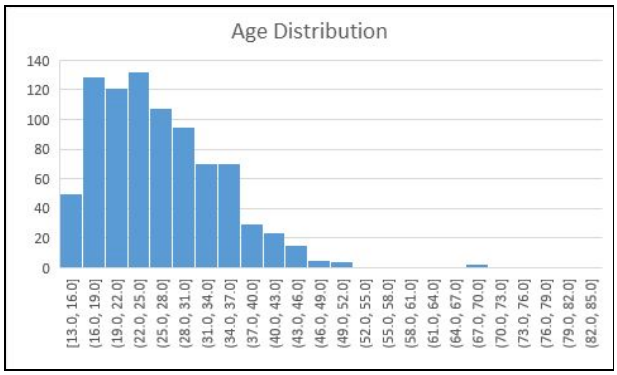
Tanmay Jha Ahaan Nachane Bakulesh Singh

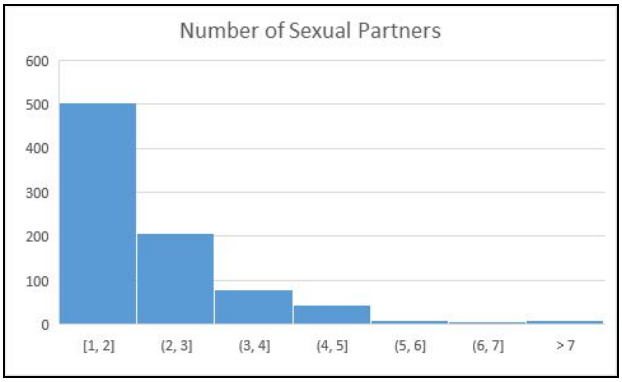
Problem Specification

Cervical cancer is the abnormal growth of cells in the cervix. The cervix is the lower part of the uterus that opens up into the vagina. Although Cervical Cancer is one of the most preventable type of cancer it kills about 4,000 women in the US and about 300,000 women globally. About 11,000 new cases of cervical cancer are diagnosed each year.
We intend to leverage a patient's medical records and behavioral factors to capture their potential risk of cervical cancer. For our project we use recommendation of a biopsy by an oncologist as the metric for potential cancer risk, because biopsy is generally recommended only when there is a significantly high perceived risk of cancer. By itself, a biopsy is an expensive and invasive procedure and we believe that a model that captures this information can be valuable a patient before they choose to undertake this expensive procedure.

Data Description

The dataset we used originally had 36 features and 858 examples of patients. 25 of the features are categorical (binary) whereas the remaining have discrete data regarding the patient. The main behavioral indicators (based on our research and a few interviews with doctors) of cervical cancer that the dataset covers are the number of pregnancies, contraceptive use, sexually transmitted diseases (STDs), smoking history, intra-uterine disease history and hinselmann test that are used by medical practitioners to recommend a biopsy. Our dataset contains a column titled Biopsy, whose value tells us whether a biopsy was recommended for this patient or not. Our goal is to predict the value of this field for new patients, since a biopsy in an expensive and an invasive procedure. Since biopsies are only recommended in cases when cancer is strongly suspected, we plan on using features which are commonly considered risk factors for Cervical Cancer to decide whether to recommend a biopsy or not.
We have included a few visualisations and some descriptive statistics to better understand the dataset





Most of the women in the dataset are under the age of 40 and the average age is a little less than 27. The average number of sexual partners that the women in the dataset have had are 2 and they range from 1 to 14. The range seems large but 97% of the women have had 6 or less partners.The average number of pregnancies in our dataset are 2 and they vary from 0 to 11. About half of the women in the dataset have had more between 2 and 4 pregnancies.
About 60% of the women in the set have reported using hormonal contraceptives. The average number of years of usage are 2 and range from 0 to 30. 60% of the women have used the contraceptives for upto 1 year.

Data Cleaning

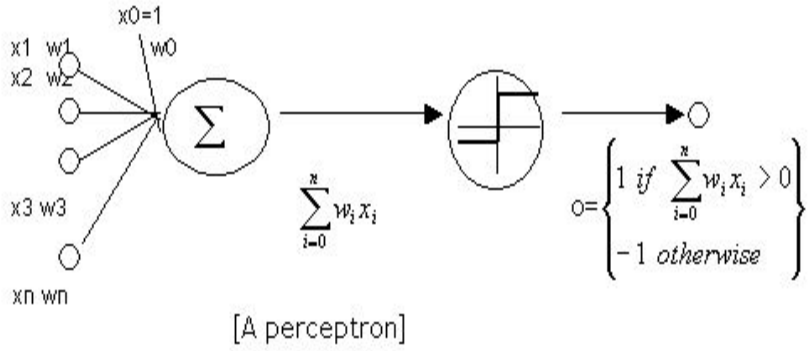
The dataset had '?' for a lot of women across attributes. We could ignore these data points because our dataset was not very large to start with. Moreover the rows of entries for which biopsies were recommended were already very less and we could not remove any of these.
We managed to clean the data by replacing the '?' with values that would result in the least bias during classification and be consistent with the labels. For example: in the number of pregnancies we divided the incorrect entries into two categories one for women who were recommended a biopsy and one for those who were not. For either of these we replaces them with the most commonly occurring entry for the classification in the remaining data. This ensured that despite having changed the entry in the field the outcome would not be affected.
We followed a similar approach for other features in order to clean the data replacing with mean, median or mode to keep it consistent with the labels and trying to add a minimal amount of bias to the data.

Model Selection and Results

We used the following models on our data:Perceptron, Perceptron with class weights, Random Forest, Random Forests with class weights, SVM with rbf Kernel, SVM with rbf Kernel and class weights
We also split our data set into cancerous and non cancerous data. The reason we did this splitting for testing is as follows - since cervical cancer is a rare disease, say 3 out of 100 patients have it. If our model is faulty, then even if it says "Does not have cancer" for every test point, it will end up being right 97 percent of the time, because only 3 people actually have cancer. To combat this, we test our accuracy separately on people with cancer and people without cancer. Further, since the incidence of cancer is so rare, our training data is heavily lopsided with non cancerous data points. In order to combat this, we used the "class weights = balanced" parameter, which is designed to account for such a lopsided training dataset.

Perceptron

The first model we decided to use on running our data was the perceptron. The perceptron algorithm is essentially a linear classifier which classifies data into two groups, which for our project is whether or not a patient is at risk for Cervical cancer. The perceptron is a form of neural networks that makes its predictions based on a linear predictor function combining a set of weights with a feature vector. It initializes these weights and sets a threshold. The perceptron training algorithm then iteratively learns according to the perceptron learning method just as we covered in class. We performed cross validation of our perceptron algorithm runs in order to check for overfitting in our model.



[A perceptron]

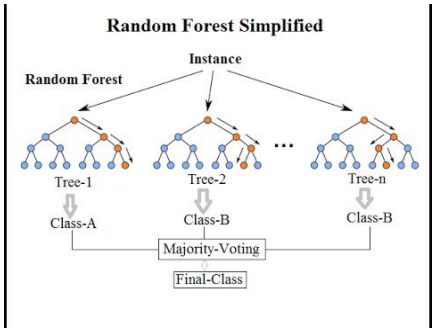
The results of our perceptron run are:

Perceptron on training data: 97%
Perceptron with Cross Validation on test data:
On Running Perceptron using Cancerous test data: 96%,
10 - fold Cross Validation using Cancerous test data: 94.5%,
Perceptron scores on Non Cancerous test data: 95.7%,
10 - fold Cross Validation on Perceptron using Non Cancerous test data: 93.19%, .

Random Forest using Decision Trees

The next model we considered is the random forest. The random forest works by creating a group of decision trees where each tree is assigned a number of features at random by using the whole training set. Each decision tree is then split on each feature with the purpose of maximizing the amount of feature we can gather from each split such that we move in the direction of maximum reduction in total entropy. Finally the decision trees are all used to take a majority vote that classifies each patient into having risk of cancer and not having a perceived risk. For the random forest model we used 100 trees in the forest with

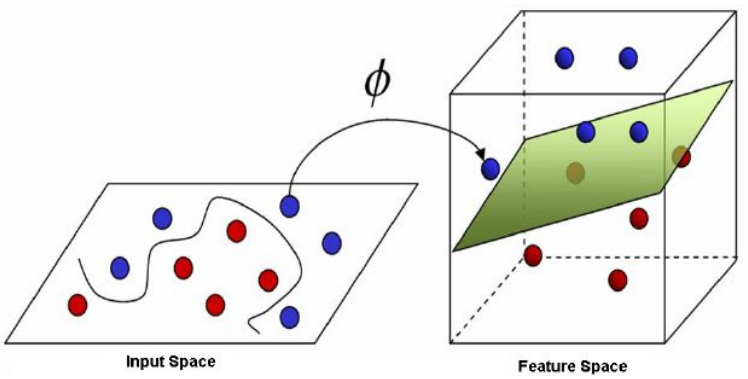
each tree to a maximum depth of 10 and the trees were permitted to randomly split n times. The random forest is not prone *to overfitting since it uses bagging and the maximum depth acts as a form of regularization that prevents a single tree from overfitting*. It does not tend to underfit since the trees can keep splitting to achieve a training error as close to zero as possible. A good illustration for random forest is as shown below:



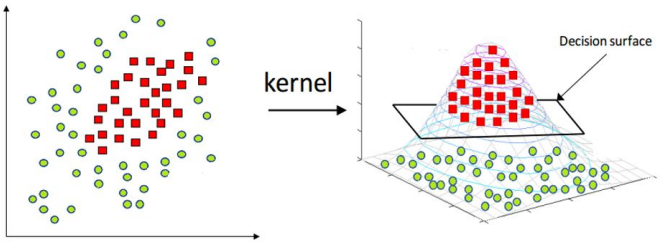
The results of random forest on our overall test set and test set split are as outlined below.
Random Forest using using test data: 96%
Cross Validation on Random Forest using test data: 95.4%
Random Forest using Cancerous Test Data: 16.32%,
10 - fold Cross Validation on Random Forest using Non Cancerous data: 95.73%
Random Forest using Non Cancerous Test Data: 99.57%,
10 - fold Cross Validation on Random Forest using Non Cancerous data: 95.73%.

Then we modified the dataset by splitting it up into 6/7ths of the cancerous data in one set and 1/7 of the non-cancerous data in one set in order to ensure consistency in proportions in every test run.

Linear SVM with rbf kernel:
Another model we chose to consider was a Linear Support Vector Machine. A Linear Support Vector Machine, like a Perceptron, separates linearly separable data by finding a decision function which uses some data points as Support Vectors. However, unlike a Perceptron, it finds the “best” classification boundary, which is the maximum margin hyperplane. However, to make this model really powerful, we used a kernelized SVM. Kernelization is a process in which data is projected to a higher dimension space, similar to Feature Engineering we studied in class, and then we run a Linear SVM on this higher dimensional new data, and project it back to its original dimension. Because of this clever feature engineering based design, we can use a Linear SVM to separate data which is not linearly separable. This can form decision boundaries in shapes of spirals, circles etc. We chose an RBF kernel, which projects data into an infinite dimensional space, and it is highly likely that it can capture almost a continuous decision boundary of any shape.



• Kernelized Support Vector Machines



Kernelized SVM with test data: 94.4%
10 fold Cross Validation on Kernelized SVM with Test Data: 93.5%
Kernelized SVM using Cancerous Test Data: 93.12%,
10 fold Cross Validation on Kernelized SVM on Cancerous Test Data: 92.15%
Kernelized SVM on Non-Cancerous Test Data: 96.42%
10 fold Cross Validation on Kernelized SVM on Non Cancerous Test Data: 95.45%
10 fold Cross Validation on Kernelized SVM Model on Proportionally mixed Dataset: 93.61%

We decided to not use K nearest neighbors, due to data sparsity. Since we have high dimensional data, the curse of dimensionality prevents us from effectively using K-nearest neighbors, unless the data lies on a manifold

Combating Overfitting:

Since our dataset was small, overfitting was a crucial concern. As described above, we used several techniques to account for overfitting

1. Class Weights - Since the incidence of cancer is so rare, our training data was heavily lopsided with non cancerous data points. In order to combat this, so that our models do not get excessively influenced by Non Cancerous training points, we used the “class weights = balanced” parameter, which is designed to account for such a lopsided training dataset. It essentially allows the model to be equally influenced by both the categories.

2. Separate Testing for cancerous and non cancerous data - as described above
3. Instead of using standalone Decision Trees, we used Random Forests, which are an ensemble method that use bagging to combat overfitting
4. We used Kernelized SVMs. These are less likely to overfit data as compared to Perceptrons and Non-Kernelized SVMs

Extension:

Even after the above measures to combat overfitting, we were getting uncannily high accuracies on our testing datasets. As an extension to our original project, we decided to investigate this. We are classifying this as a separate section since it ended up being a mini-project on its own.

When we started this project, we wanted to capture a patient's risk for Cervical Cancer based on their medical records and behavioral factors. We chose to predict biopsy as the best metric for capturing this risk.

Since we were getting high accuracies and suspected overfitting, we tried to find more data. After extensive searching, we could not find this data online, so we decided to get more data by getting new data points labelled by an oncologist. We reached out to a practising Oncologist to get this new data, but we gained some crucial insights which motivated us to think of other approaches instead of using biopsies as the metric for measuring cancer risk.

After speaking with him we realised that a biopsy is only recommended in cases the screening tests, namely Hinselmann, Schiller and Cytology come back positive. Since our dataset already had these three tests represented as features, this was skewing the accuracy results, since according to the physician, the recommendation for getting these tests done is given only if there is a high risk of cancer based on other behavioural factors and medical data, and if these tests are positive, a biopsy is almost always recommended. These tests are essentially the same with some differences in their methodology and are recommended based on availability of testing facilities.

Therefore, we concluded two things:

1) Since the presence of these features has a very high correlation with recommendation of a biopsy, and these tests are a part of the medical pipeline which results in a biopsy recommendation, we could gather interesting insights if we were to remove these from our dataset in order to capture risk for cancer based on behavioral factors and other medical data. **Then we would able to predict risk for patients who haven't yet been to doctors, and don't already suspect cancer incidence.**
2) Since these tests are usually recommended if there is a significant perceived risk for cancer, based on a Doctor's assessment, we might be able to capture the risk more accurately by trying to predict these three labels.

So we decided to remove these tests as features and instead use them as three separate labels for behavioral risks associated with cervical cancer. Then we can train three separate models with each of these labels. Once we have this model we can predict a behavioral and medical risk if even one of these models returns a value of 1 (meaning the test was recommended because the physician perceived a behavioral pattern associated with risk of cervical cancer).

Due to the lack of resources and time we could only get in touch with one oncologist and we got a set of 29 points which we used for additional testing.

Model Selection for Extension

For our extension, we decided that since predicting that someone does not need to go to a doctor can have serious ramifications if the person ends up actually having cancer, we need probabilities of how confident we are in our predictions. Therefore, we decided to use Logistic Regression, since that outputs probabilities of our confidence in our predictions.

We proceeded to use Logistic Regression to predict the new labels of Hinselmann test, Cytology and Schiller's test. We used a logistic classifier so that we could also retrieve the probability associated with the three labels in each case. Thus we would have a 3X2 matrix of probabilities associated with each of the two labels (0 and 1) for each of the three tests. We would only recommend not going to a doctor if the probability associated with 1 is less than a very conservative threshold of say 10% for each of the three tests. We got the following results for accuracy.

score_cv_log_citology	float64	1	0.75119412124923435
score_cv_log_hinselmann	float64	1	0.97999941679059877
score_cv_log_schiller	float64	1	0.65852214737701575

Since these labels represent the behavioral risk factors, predicting them gives us a reasonable estimate of the risk a particular individual faces.
In order to make our model more resilient, we decided to do feature selection.

Feature Selection for Extension:

We considered several ways of selecting such as:

1. Low Variance Features - This method identifies features that have a very low variance and eliminates them because they could not possibly be cause a change in the labels. However, this would not make a lot of sense for us because most of the features we have are binary in nature.
2. Univariate Tests (like the Chi-square): Such tests usually depend on assumptions of NIID effects of the various features and are usually used to provide an estimator of feature importance for the two methodologies that follow. Since it essentially boils down to a choice between using an external distribution (based on the assumptions mentioned above) for assessing the importance of

a feature and the loss function we would eventually be using to create the model we decided to go with loss function since we were not very confident that our dataset would adheres to these assumptions.

3. RFE_CV - In this test, recursive feature elimination is done with cross validated losses in order to select only the features that have the biggest net impact on the loss function. Therefore this method gives us the most important features. When we ran this test on our original dataset (which contained the features for Schillers, Heinselman and cytology tests) with linear models like Perceptron, SGD, Linear SVC and Logistic regression, all of them selected the Schiller's test as the most important feature. **Thus, this confirmed what the physician told us - that physicians recommend biopsies based mainly on the results of these tests.** After we removed these features, we ran RFE_CV with a Logistic Regression model, and found HPV and CIN to be the most important risk factors, which again confirmed and followed the physician's recommendations. This validated our models initial performance for us.
4. RFE - This test is very similar to RFE CV. However, in this case, we specify the number of features we finally want, and the RFE test selects that many most important features for us. It removes the "less important" features and keeps the features with the most impact on the loss. We asked RFE to select the 25 most important features for each of the 3 labels - Hinselmann, Cytology and Schiller's test.

Training and Testing on the Smaller Dataset:

We finally used Logistic Regression on this new dataset of 25 features to train and predict the labels for each of the 3 labels. Then we tested these models on both our testing data from our pre-existing dataset and from the new data we obtained from the physiologist. We got the following results.

score_log_test_citology_mod_feat	float64	1	0.75159235668789814
score_log_test_citology_new_data	float64	1	0.62068965517241381
score_log_test_hinselmann_mod_feat	float64	1	0.72611464968152861
score_log_test_schiller_mod_feat	float64	1	0.71974522292993626
score_log_test_schiller_new_data	float64	1	0.7931034482758621
score_log_hinselmann_new_data	float64	1	0.55172413793103448

Promise for Commercial Application:

Using very conservative probability threshold for each of the three tests in conjunction this test can be used to assess cervical cancer risk associated with behavioral factors. The models used on the entire dataset using a patient's entire medical records provide a very effective model that can be used to suggest whether or not a patient is at risk for cancer. However, we are cautioned by the fact that the model has been trained and run on a relatively small dataset so in order to scale this to a commercially viable product

we would need to source more data from patients and oncologists as we did in the project, but on a larger scale so that the model can be scaled effectively

Our extension considers the patient's medical records without the data for these 3 tests and only containing behavioral factors and partial patient history. The model trained on only these features, while it provides a lot of good insights, by itself may not be a commercially viable option. But both the primary model and the extension in conjunction with each other could act as a good indicator for a patient's risk for cancer.

Future Studies:

We would recommend that given the time and resources this strategy be adopted to build an even better model by consulting with a panel of physicians and getting a large dataset from them. This set would be better because it comes from not one but several physicians therefore ensuring that it is IID.

After this the same strategy as adopted by us can be used to build a model for each of the three tests and the resulting probabilities can be used to make recommendations as explained above.