DSC 630 Project Proposal

Eric Dickey, Emily Phillips, Jesse Zamora

Department of Data Science, Bellevue University

DSC 630: Predictive Analytics

Professor Fadi Alsaleem

December 10, 2021

**Introduction**

There are many significant events in life which people may think they would never be able to predict or gain awareness of; one of these being the onset of various diseases, cancers, etc. However, with the power of predictive analytics, especially in the medical field, this mystery does not seem so dooming! Strides are being made in being able to predict whether a tumor is malignant or benign, which symptoms or side effects can predict illnesses, and so on. Through this project, the focus will be around insights on breast cancer.

According to the American Cancer Society, “the average risk of a woman in the United States developing breast cancer sometime in her life is about 13%. This means there is a 1 in 8 chance she will develop breast cancer”. Additionally, breast cancer is the second leading cause of cancer death in women (2021). However, due to women finding breast cancer earlier in recent years, the breast cancer death rates in women have begun to decline. Early identification is directly related to proper screening, which comes along with being able to accurately predict breast cancer tissue images’ diagnoses.

The predictive analytics in this project will utilize past breast screening tissue image data to classify future images’ diagnoses as either benign or malignant. A benign diagnosis refers to a diagnosis that the breast tissue is non-cancerous, whereas a malignant diagnosis would specify that the tissue is harmful and cancerous. Through selection of appropriate models and model specifications, the goal and scope are to identify a model which has a high accuracy in identifying tissue images of concern, which can be extremely beneficial for early identification and quick intervention for the longevity of patients’ lives.

Furthermore, in scope, the team will be working on this project throughout the next twelve weeks. Python and R programming languages will be utilized to explore breast cancer tissue imaging data and to transform any fields as needed for model applications and evaluations. The project management plan is specified towards the end of this proposal for more clarity around the team dynamic and approach. The rest of this project proposal will dive into the project’s technical approach for exploratory data analysis and modeling application, which will eventually allow for uncovering data-driven decision making around the problem statement. Background will also be given around the data sources for the breast cancer tissue images, and a plan will be specified for testing and evaluating the chosen models. Along with these initial requirements, the benefits and risks of model deployment will also be discussed.

**Data Source**

The dataset chosen for this project was sourced from the University of California Irvine’s Machine Learning Repository. Considering the source of this data originates from a respected educational research institute, there are limited concerns regarding the integrity or quality of the data. The dataset contains features that were extracted from a digitized medical image used to form a breast tissue sample and has a defined target variable that classifies the data points as malignant or benign. The variables within the dataset contain various measurements related to the cell structure, including the following for each cell nucleus:

|  |  |
| --- | --- |
| * Radius (mean of distances from center to points on the perimeter) * Texture (standard deviation of gray-scale values) * Perimeter * Area * Smoothness (local variation in radius lengths) | * Compactness (perimeter^2 / area - 1.0) * Concavity (severity of concave portions of the contour) * Concave points (number of concave portions of the contour) * Symmetry * Fractal dimension ("coastline approximation" |

The mean, standard error, and worst (mean of three largest values) were computed for each image resulting in a total of 30 features. This data set is suitable for a supervised machine learning project as it has a clear target variable defined, which is the diagnostic value of malignant or benign.

**Technical Approach and Analysis**

The technical approach the team will use to create a model using the breast cancer dataset will be structured in a manner that is logical and leverages the information obtained or created from the previous step. The first step in building a model to assist in predicting breast cancer diagnoses is to collect the data. The data is available in a flat file formatted as a CSV.

As a part of preparing the data for analysis and conducting exploratory data analysis on the dataset, R will be leveraged to explore the data. The data will be evaluated to determine if any data inconsistencies need to be addressed, such as inconsistent values of variables, null values, and outliers. The data will also be visualized to check if there are correlations between the values, which will be important because it appears that there are similar variables contained within the dataset. Distribution will also be visualized, and the balance of target variables will be explored to help prepare the data for training the model.

Once the exploratory data analysis has been completed with R, the data will be moved to a data frame within Python for the duration of the work to build a model. After determining the correlation between the variables in the data set, feature engineering with help to reduce the number of features used in the final model to make the model less complex and improve accuracy by removing highly correlated variables. Feature importance will be calculated using a Random Forest Model, which provides good predictive performance in classification models. The importance will be determined by calculating how much each variable contributes to the decision on a decision tree using Gini Importance. Stack ranking the feature importance will help to narrow the number of features used to evaluate various model performance.

**Testing and Evaluation**

Following feature engineering, several models appropriate for classification will be evaluated for performance. Model performance will be determined by comparing the accuracy of several models. With the best performing model identified, the parameters will be tuned by adjusting the number of iterations over the training data. The performance will be evaluated further by adjusting the number of features until the optimal parameters are identified to yield the greatest accuracy.

**Model Requirements and Deployment**

The model will be designed to ingest a flat file containing digitized imaging data. The requirements for the model include a storage solution for the flat file. While the data does pertain to medical related data, the data is anonymized and does not appear to contain any personal identifiable information that would need to adhere to HIPPA guidelines. The model itself will need to have access to an environment with Python 3.6+ and will eventually need to be scaled with compute power to be able to handle large amounts of data. A cloud computing environment appears to be designed to address this need and to offer increased processing power as the model grows and ingests more data.

**Expected Results and Broader Deployment**

A model that has a specific medical imaging use case can bring tremendous value to diagnostic imaging and help achieve a faster and more efficient diagnosis that could improve the survival rate for patients. With the rails for the model built and specific data processing related to the cell structure for biopsies completed, it should be explored if the model can be expanded to other use cases which evaluate the similar features to the model, including cell structure measurements. This model could serve as the framework for broader deployment. Additional oncology diagnostic screening should be evaluated to determine if there are other tests which use the same or similar features and evaluate if the model’s performance is within an acceptable accuracy to expand its use. This could lead to potential rapid deployment of the model if it can be used in a manner that is agnostic to the location of cancer but evaluates the cell structure the same.

There are also additional features that could be added to a model to potentially increase the accuracy of the model even further. Additional variables, such as geographic region, exposure to cancer-causing environments, family history, and lifestyle factors which increase the risk for cancer could all be relevant features to explore. There could be weights applied in scoring the inputs which carry an inherently higher risk of cancer due to those additional factors. This would allow the model to potentially be more aggressive in scoring and flag an image as a higher risk for cancer which would require a manual review or additional testing. As the model matures and is accessible to the larger medical community, a REST API integration can be used to allow for remote interactions with the model, including ingesting data and returning the results.

**Execution and Management of Project**

In organizing the work for the project, an agile-like methodology was adopted. With agile development, iterative development is encouraged, which facilitates continuous improvement and continuous deployment. A feedback loop is integral to the team’s functioning, and throughout the project work, consistent and clear communication with one another will be maintained, so everyone is always on the same page.

The team will meet weekly and utilize the time to distribute tasks and assign estimations for work completion. In doing so, progress will be continuously updated, and any blockers can be identified and discussed. For any blocked work, this will require full cooperation and coordination to understand where help is needed and when/how it can be given. The standard for work completion is that it will be completed and shared amongst the team no later than Saturday, in order to allow for review and team acceptance by the due date submission on Sunday nights.

Also, in terms of having a feedback loop for technical coding, a shared GitHub repository has been created, that will also have controls to avoid code overrides and disruption of the master branch. The focus as a team in terms of executing and managing the project is to maintain and uphold clear communication with one another, while also participating to the fullest capacity (as able). The promise is to deliver valuable results and to be able to implement a project which drives insights around the decided problem statement. Teamwork, self-organization and accountability are key concepts to the project plan, and hopefully with these in mind, project risk can be mitigated as much as possible.

**References**

*How common is breast cancer?: Breast cancer statistics*. American Cancer Society. (n.d.). Retrieved December 12, 2021, from <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>.