C964 Computer Science Capstone

Alon Robinson

Western Governors University

Letter of Transmittal

March 5th, 2022

Neal Bordreaux

Chief Medical Officer

Data Doctors, Inc.

Dear Neal,

Based on our previous conversations in regards to solutions our clinics can use to better serve patients, I want to inform you of a patient-focused solution for the early prediction of risk of diabetes.

As you well know, diabetes is the 7th leading cause of death in the United States (Centers for Disease Control and Prevention, 2022). It is a dangerous disease that, left undiagnosed, can cause damage to vital organs and lead to death. The early detection of diabetes risk is thus extremely important for the health and wellbeing of patients at the clinics we serve.

For these reasons, I am proposing a solution that will enable doctors to easily determine patient risk for diabetes during routine checkups. I will lead a team that will create a machine learning algorithm that can accurately predict whether a patient is at risk for diabetes, with an easy to use GUI that doctors can use. The use of machine learning will aid in preventing human error in

diagnosis, and allow the clinic to better serve patients by treating this deadly disease before it can take root.

To do this, I am estimating a timeline of 5 weeks and a total cost of $39,000. This includes one data engineer, one software engineer, and I will serve as the project manager. Both engineers graduated at the top of their class and have a combined 20 years of experience in their respective fields. They have worked for Data Doctors, Inc. 5 years and 7 years, respectively, and have proven their skills time and again through the seamless delivery of products over their tenure. They have both lectured at Harvard University and speak regularly at conferences. I know that you will find their expertise and ability to deliver this application exceeds your expectations and brings much value to the Data Doctors, Inc.

If you have any questions, please do not hesitate to contact me. We would love to get started on this game changing project as soon as possible.

Sincerely,

Alon Robinson

Director of Engineering

Data Doctors, Inc.

Project Recommendations Summary

Problem Summary

This proposal describes the need for a machine learning model in order to address the needs of Data Doctors, Inc. Data Doctors, Inc. owns and operates several healthcare clinics across the city of New Orleans and strives to combine human and machine expertise in order to better serve and address patient needs. From 1999 to 2014, the prevalence of diabetes in Louisiana among those ages 50 to 64 jumped from 11.5 percent to 17.8 percent (Ted Griggs, 2016). When a person already has diabetes, it is easy enough to diagnose with a blood test, but being able to catch early signs of oncoming diabetes will enable better treatment for patients.

Application Benefits

Data Doctors, Inc. will use Logistic Regression, a type of machine learning algorithm used for classification. For our purposes, we will have two categories: at risk and not at risk. With the use of machine learning, the data we collect from patients will be quickly analyzed and a prediction given. The benefits of this is that the trained machine learning model is able to spot trends in the data it was trained on that were associated with diabetes diagnosis, and use these insights to predict patient risk at a much higher accuracy than a single doctor. With this added insight, doctors are more able to care for patients early and prevent diagnosis, rather than treat it.

Application Description

The application that the doctors will interface with is a web based application hosted on Heroku. The application will take numerical data inputs that correspond with each patient, and display a message stating whether a patient is at risk of diabetes or not. Doctors will have a dedicated login in order to access the application. The application was built using Python 3 and Bottle, a

microframework used for creating web servers. The interface was created with HTML, CSS, and JavaScript. There is no database needed as the results are computed on the fly. With this interface, the doctors will have easy access to enter the patient input and receive immediate results. This will enable quick diagnosis and better patient care.

Data Description

The data used to create this model is originally from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This data tracked pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function (a computation of likelihood of diabetes based on family history), and age of a group of Pima Indians. For each of the women in the dataset, there is also a classification of whether or not they had diabetes. This was then used to train our machine learning model using Logistic Regression. This dataset is publicly available and has been widely studied, so there are no concerns with our use of it in this model.

Objective and Hypothesis

The objective of this project is to detect early warning signs of patients who are at risk of diabetes through the use of machine learning. We believe that our trained model will be able to spot trends that can lead to diabetes much better than a human, and will be able to aid in the categorization of patients who need to be treated before they become diagnosed with diabetes. As a result, we believe that this model out of the gate will be able to catch 60% more patients at risk for diabetes than the clinics normally do without machine intervention. The model will be 80% accurate in its predictions.

Methodology

Development of the machine learning algorithm will follow the SEMMA methodology. SEMMA methodology was developed by the SAS Institute to describe the methods used to gain insight via analysis of datasets. Since this project is heavily focused on analyzing data and making predictions based off of it, we feel that this is the best methodology for building the algorithm. Below you will find an outline of how we will implement this methodology for the project.

- Sample: We will retrieve the Pima Indian diabetes data from NIDDK
- Explore: With the variables and factors found, we will begin analyzing the data in order to find patterns, relationships and any gaps in the data.
- Modify: We will take the data that we have and parse it and clean it up (accounting for gaps, etc.) in order to be able to model it.
- Model: We will apply different data mining techniques to the data in order to efficiently model it for our purposes.
- Assess: This is where we evaluate the model and are able to see how it performs on the data we have.

Funding Requirements

Resource	Description	Cost
Software Engineer	Will implement findings from data scientist	$75/hr

Data Engineer	Will look at data and determine what insights can be made and how	$75/hr
Product Manager	Will ensure the project is run smoothly and write up final data report	$45/hr
	Total	$195/hr * 200hrs (5 weeks) = $39,000

Stakeholders Impact

The biggest stakeholder of this project is the Chief Medical Officer of Data Doctors, Inc. He works closely with the engineering department to develop all of the products used in the clinics and ensure their accuracy and ease of use for the doctors and nurses on staff. Because of the high prevalence of diabetes in the community, the Mayor of New Orleans has agreed to give a reimbursement grant of the costs of development after we are able to prove the application's effectiveness. The Mayor has a vested interest in seeing the communities in New Orleans become healthier.

Data Precautions

The data used for training the machine learning model is publicly available. The data collected from patients that will be input into the machine learning algorithm does not have any identification data, and does not store any results. As a result, there are no legal or ethical considerations with this particular project.

Developer's Expertise

The Software Engineer and Data Engineer both graduated at the top of their class and have a combined 20 years of experience in their respective fields. They have worked for Data Doctors, Inc. 5 years and 7 years, respectively, and have proven their skills time and again through the seamless delivery of products over their tenure. They have both lectured at Harvard University and speak regularly at conferences. The Product Manager has 15 years of experience bringing large product offerings to completion, having worked with small teams of 1 engineer up to large teams of 100 engineers. In addition, all 3 have worked together on long term projects in the past. As a result, they have a level of working chemistry already and won't have to wade through the awkward phase of working relationships. I am quite certain that our skills, communication and ability to bring this project to completion are unmatched.

Project Proposal for IT Professionals

Problem Statement

As the largest diabetes center in the city of New Orleans, Data Doctors, Inc. is faced with the treatment of many patients who come in. After getting the lab work for each patient, the process of evaluating the results and determining a risk of diabetes takes a good bit of time and effort. In addition, each doctor is seeing a minimum of 20 patients per day, which can result in human error due to the amount of information that needs to be kept up with. Data Doctors, Inc. is looking for a solution that will provide a predictable interface for data input, and give extremely fast and accurate results. Data Doctors, Inc. wants a solution that will eliminate the human error and time consuming effort it takes for patients to receive a diagnosis, thus enabling them to focus more on patient care.

Customer Summary

This product will be used by the doctors and nurses at Data Doctors, Inc.. It will replace the manual analysis done by doctors who often see 20 or more patients per day, and move the burden to the machine learning algorithm. The algorithm will be deployed to the cloud and the doctors given the url and login information. Because the application is simple to use, the engineers will take 1 day to train the doctors and nurses how to use it and leave instructions.

Existing System Analysis

There are currently no systems or data products in place that aid in the prediction of diabetes risk at Data Doctors, Inc.. As a result, they currently rely on manual analysis of patient lab work. These calculations can be tainted by human error based on the amount of patients each doctor sees in one day, due to the amount of information that they must keep track of. Following the

completion of the diabetes prediction application, there will be no need for manual analysis, as the machine learning algorithm will take on the burden of classifying whether or not a patient is at risk of diabetes or not.

Data

The data used to create this model is originally from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This data tracked pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function (a computation of likelihood of diabetes based on family history), and age of a group of Pima Indians. For each of the women in the dataset, there is also a classification of whether or not they had diabetes. This was then used to train our machine learning model using Logistic Regression. This dataset is publicly available and has been widely studied, so there are no concerns with our use of it in this model.

Project Methodology

Development of the machine learning algorithm will follow the SEMMA methodology. SEMMA methodology was developed by the SAS Institute to describe the methods used to gain insight via analysis of datasets. Since this project is heavily focused on analyzing data and making predictions based off of it, we feel that this is the best methodology for building the algorithm. Below you will find an outline of how we will implement this methodology for the project.

- Sample: We will retrieve the Pima Indian diabetes data from NIDDK

• Explore: With the variables and factors found, we will begin analyzing the data in order to find patterns, relationships and any gaps in the data.

• Modify: We will take the data that we have and parse it and clean it up (accounting for gaps, etc.) in order to be able to model it.

• Model: We will apply different data mining techniques to the data in order to efficiently model it for our purposes.

• Assess: This is where we evaluate the model and are able to see how it performs on the data we have.

Project Outcomes

Following the completion of this project, the doctors at Data Doctors, Inc. will be left with a functional GUI that can be used to predict risk of diabetes in patients. They will have the ability to input data from patients' lab work, and get an immediate analysis of patient risk of diabetes. They will also receive ongoing support over the 1 year following completion to track the accuracy of the predictions and fine tune the algorithm if necessary. This application will be deployed to the cloud for ease of use, with the url and login access given to the doctors.

Implementation Plan

This project will be implemented over the course of 4 weeks, with 1 week for each of the first 3 phases of the SEMMA model, and the last 2 phases combined in the final week. The team will take part in short, daily standups with stakeholders in order to keep them updated on the status

of the project. At the end of each sprint, until the data project is beginning to become a tangible, interactive project, the team will provide a recap of all that has been worked on, along with current findings from the data. Once the GUI is being implemented, the team will provide demos of the application until it is completed.

Evaluation Plan

This algorithm will be evaluated in two ways. The first will be comparing the classifications of patients done by the machine learning algorithm to manual analysis of those same patients by the doctors. This will give us a decent, but nowhere near perfect, indication of the algorithm quality. We expect the algorithm to catch more patients at risk than the doctors. The second way will be to take a look at previous patients' lab work that have already been classified by the doctors, feed those data points into the machine learning algorithm and see how it performs. We will use patients who were at risk or positive for diabetes, along with patients not at risk. If the algorithm classifies patients who were labeled not at risk previously by the doctors, they will be asked to come back in for reevaluation. Our team will offer long term support for 1 year following the project completion in order to ensure accuracy and fine tune the algorithm if needed. After this period, the application should be in stable condition because of its’ minimal size.

Resources and Costs

The application will be a desktop application that can be installed and run on Windows or Mac. We recommend having the latest versions, as that is what we will be using for development.

However, the tools we are using to build the application can be targeted for older versions if necessary. The application, its licensing and usage, and 1 year of support is included in the project costs. Additional support can be purchased that will be billed at a rate of $100/hour.

Below is a table with the total project cost:

Resource	**Description**	**Cost**
Software Engineer	Will implement findings from data scientist	$75/hr
Data Engineer	Will look at data and determine what insights can be made and how	$75/hr
Product Manager	Will ensure the project is run smoothly and write up final data report	$45/hr
	Total	$195/hr * 200hrs (5 weeks) = $39,000

Timeline and Milestones

Event	Start Date	End Date	Developer hours required	Dependencies	Assigned Resources
Project Start	June 6th, 2022	June 10th, 2022	0	None	Project Manager, Stakeholders
Phase One (Sample)	June 6th, 2022	June 10th, 2022	40	Project Start	Project Manager, Developers
Phase Two (Explore)	June 13th, 2022	June 17th, 2022	40	Phase One	Project Manager, Developers
Phase Three (Modify)	June 20th, 2022	June 24th, 2022	40	Phase Two	Project Manager, Developers
Phase Four (Model)	June 27th, 2022	July 1st, 2022	40	Phase Three	Project Manager, Developers
Phase Five (Assess)	July 4th, 2022	July 8th, 2023	40	Phase Four	Developers, Stakeholders

Support	July 11th, 2022	July 11th, 2023	Ongoing developer support	Phase Five	Developers, Stakeholder

Post Implementation Report

Project Purpose

Data Doctors, Inc. is the largest diabetes center in the city of New Orleans, and sees many patients. They needed a way to easily and accurately analyze the results of patient lab work in order to assess risk of diabetes. Currently, the doctors who work at the center see 20 or more patients per day. This can contribute to human error in the analysis of each patient's lab work, due to the amount of patient data each doctor is tasked with handling. In addition, the analysis needed to determine a patients' risk for diabetes is time consuming and results in less time spent caring for patients. With the machine learning solution provided, Data Doctors, Inc. has been able to quickly get analyses of a patient's risk of diabetes that removes the human error. We provided a web based GUI that takes input data for each patient and returns a result letting the doctor know if a patient is at risk or not.

Datasets

The data used to train the model came from the NIDDK. This data was analyzed and found to be an already clean dataset with no null values. Below is an example of the data.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Data Product Code

The machine learning model was built using Python, Numpy, Pandas, and scikit-learn. We were able to use Pandas to create charts and plots of the data to use for analysis and draw conclusions. We found that the biggest correlation with diabetes outcome is the glucose level. With these insights, we used Logistic Regression to train the model, pickled it, and imported it into our web application. The source code can be found at https://github.com/algoasi/capstone.

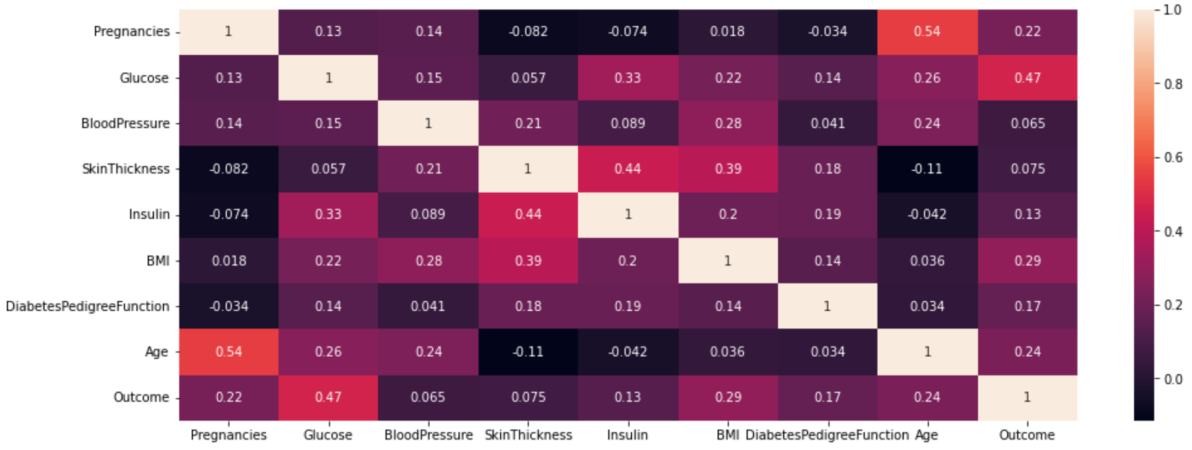
Hypothesis verification

The immediately verifiable hypothesis was that the model would be 80% accurate in its predictions. This hypothesis was missed as the accuracy of the machine learning model fell just shy at over 78% accuracy. Through the 1 year of ongoing support, the team will be able to adjust the model to meet higher accuracy.

The other hypothesis is long term, which is that the machine learning model would enable the clinic to catch 60% more patients who are at risk of diabetes. With the amount of patients that come into the clinic, we believe the benefits of the application will be seen quickly.

Effective Visualizations and Reporting

Before training the model, it was imperative that we take a closer look at the data. With the use of Pandas, we were able to find the strongest factors that influenced the outcome. The biggest influence on outcome was glucose levels.



In the above heatmap, we plotted the correlations of the other values as they relate to Outcome. Glucose had the highest correlation with a rating of 0.47, followed by BMI with a rating of 0.29, and age with a rating of 0.24. This means that women with higher glucose, BMI, and age were more likely to have diabetes than women who were lower in those values.

Another correlation we can see from this heatmap is that Insulin levels have a not insignificant correlation with Glucose. In addition it has an even higher correlation with skin thickness. But in even though Glucose has a high correlation with diabetes Outcome, we don't see much correlation between skin thickness or Insulin levels and diabetes Outcome.

We can also see that age has some association with Glucose levels.

Armed with this information, the doctors at Data Doctors, Inc. will be better able to spot potential risks for diabetes in patients

Accuracy analysis

The machine learning model has an accuracy score of 80.21%, meaning that it correctly classified whether patients in the data set did or did not have diabetes 80.21% of the time. We also measured precision, recall, and F1 score in order to get the best picture of how the model performed.

	precision	recall	f1-score	support
0	0.82	0.91	0.86	130
1	0.75	0.58	0.65	62
accuracy			0.80	192
macro avg	0.78	0.74	0.76	192
weighted avg	0.80	0.80	0.79	192

This classification report can be found inside the accompanying Jupyter notebook

Application Testing

The application itself is very simple and, as such, was easy to test. It consists of a webpage that takes input that is submitted to an api endpoint. We were able to confirm that the application accepted all inputs. We also confirmed that the data was being sent to and returned from the API endpoint by examining the network requests in the Chrome developer tools panel.

Application Files

The application is deployed to Heroku and can be accessed at https://algoasi-capstone.herokuapp.com/index.html. If you want to run the application locally, you can find the complete code repository at https://github.com/algoasi/capstone. The code repository has a detailed README file that outlines how to run the application, and includes the accompanying Jupyter notebook.

User's Guide

A detailed user guide can be found in the Github repository for this application https://github.com/algoasi/capstone. In the README, all the steps for using the application are listed out.

Summation of Learning Experience

My prior experience consists of 6.5 years of professional Web Development, so knowing how to code has been a strength in completing this project. This enabled me to put together the simple web application that hosts the machine learning program very quickly. It also assisted in learning the coding syntax for developing the machine learning algorithm.

What I had to learn, however, was machine learning itself. I had never ventured into developing machine learning applications before this project. I was able to gain understanding through accessing the WGU Udemy site and going through a machine learning and data science course.

This experience has given me valuable knowledge and experience that I can continue to build upon as my career advances. I plan to get a Masters degree in Computer Science and having some basic knowledge of machine learning algorithms and how to build basic programs will aid me in obtaining my degree.

Citations

Centers for Disease Control and Prevention. *What Is Diabetes?* 2 Mar. 2022, Retrieved from cdc.gov/diabetes/basics/diabetes.html

Ted Griggs. “Louisiana, ‘the Obesity-Diabetes Heartland of America,’ Still Ranks as Least Healthy State for Seniors.” *NOLA.com*, 25 May 2016, Retrieved from nola.com/news/business/article_b9c77f06-20c3-5bf1-9172-b7645315499a.html