June 10, 2024

C964: Computer Science Capstone Template

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 2](#_Toc168842679)

[Part B: Project Proposal Plan 3](#_Toc168842680)

[Project Summary 3](#_Toc168842681)

[Data Summary 4](#_Toc168842682)

[Implementation 4](#_Toc168842683)

[Timeline 5](#_Toc168842684)

[Evaluation Plan 5](#_Toc168842685)

[Resources and Costs 6](#_Toc168842686)

[Part C: Application 7](#_Toc168842687)

[Part D: Post-implementation Report 7](#_Toc168842688)

[Solution Summary 7](#_Toc168842689)

[Data Summary 8](#_Toc168842690)

[Machine Learning 9](#_Toc168842691)

[Validation 9](#_Toc168842692)

[Visualizations 10](#_Toc168842693)

[User Guide 12](#_Toc168842694)

[Reference Page 16](#_Toc168842695)

# Part A: Letter of Transmittal

06/07/2024

John Q. Public

DFW Insurance

123 Main Street, Dallas, Texas 76022

Dear John Q. Public,

The insurance marketplace is becoming increasingly competitive every year due to many new competitors and rapid increases in healthcare costs by many healthcare providers. DFW Insurance makes a profit by making sure the monthly premiums on average are more than the total future costs of a customer’s healthcare bills. In the past, since we were a smaller company and wanted to gain as much business and customers as possible. The company would have its agents sign up any person at the same premiums based on market pricing. They would figure out these market prices by letting the customers give them quotes from larger healthcare insurance companies and making their healthcare premiums a little lower than our competitors. This strategy has been successfully, however, every year this has become a more arduous task as healthcare costs keep increasing and we are seeing more losses as our premiums are not priced correctly. The accounting department is worried with the rising costs of healthcare if there is a more efficient way to be able to set the insurance premiums to make our company more profitable.

My proposal is that we take internal data from our previous customers and produce a model that will predict future healthcare costs of potential customers. By taking sample data points such as age, gender, BMI, number of children, smoker, and region of the United States they reside in we can produce a fairly accurate model that will help our agents set premiums for customers. This will in turn make us more profitable by gaining more customers since we will be able to attract lower premiums for lower risk customers and charge more for higher risk customers. Also, our agents will have more time for prospecting and gaining new business rather than spending time with customers trying to figure out what other insurance companies are charging.

Currently, our team of software engineers are busy with other projects keen to our mission goals. But my knowledge in machine learning, python and with help of open-source machine learning algorithms I can quickly build a model that our insurance agents can use in the field. I think in a relatively brief time, a month, and with minimum costs and resources we can have a model up and running that agents should be able to use in the field. I will use the knowledge that I gained from my Bachelor of Computer Science at WGU to help facilitate in building this application.

Sincerely,

James Powell

James Powell, Software Engineer

# Part B: Project Proposal Plan

## Project Summary

DFW Insurance is a health insurance company that charges its customers a certain premium and in return they will cover all healthcare costs that the person will incur. DFW Insurance makes a profit by making sure the monthly premiums on average are more than the total costs of a customer’s healthcare bill. In the past, since it was a smaller company and wanted to gain as many customers and as much business as possible. The company would have its agents sign up any person, at the same premiums based on market pricing. They would figure out these market prices by letting the customers give them quotes from larger healthcare insurance companies and making their healthcare premiums a little lower than their competition. This strategy has been successfully, however, every year this has become a more challenging task as healthcare costs keep increasing. Management is worried with the rising costs of healthcare if there is a better way to be able to set premiums and to set the insurance premiums. Another concern management has is agents spend a lot of time with customers trying to determine what their prices should be instead of more productive tasks. If agents had a simple tool where they could plug in data about a customer, and it could accurately predict what their total healthcare costs would be. Our agents can determine what premium they would need to pay for DFW Insurance to be profitable. The internal engineering department will develop a tool that can predict the healthcare costs based on a short questionnaire that the customers can fill out. This questionnaire can be implemented by the agent and a user guide will be created that the agent can follow along with the customer and should give the agent a relatively accurate healthcare cost prediction that the agent can in turn use to determine what price their health premium should be at. With this tool, insurance agents will be able to produce better priced premiums which should make DFW Insurance more profitable.

## Data Summary

DFW Insurance is not new to the insurance business and has a wide list of customers. Management can provide us with a list of customers that we can use to test and train our models on. But, due to increasing privacy laws and security over customer’s personal data. Management will have to send out a notice to customers’ requesting if we can use their personal data, to provide more accurate pricing. Their names will not be listed on the data and only one employee will have access to the full data and will need to scrub it of any personal information such as their name or address. The data points that management has approved to be left on will be the gender of the policyholder, the age of the policyholder, if the policyholder is a smoker or non-smoker, part of the region of the United States they reside, the number of children of the policyholder, the body-mass index of the policyholder, and total healthcare costs DFW Insurance has incurred from the policyholder. The employee will scrub all personal information leaving only the information that was listed above on an excel sheet. This will protect DFW Insurance and its’ customers from any legal liability in case this spreadsheet is involved with a data breach. This data will be needed to complete the project as we need those data points to determine the customer’s estimated healthcare costs.

## Implementation

For this project, management wants a Waterfall methodology used because it will create a predictable timeline and budget and because there are clear and stable requirements required for this project. This project will be divided into a five-step plan which consists of requirements gathering, data collection and processing, model development, model evaluation, and deployment of the application.

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Requirements Gathering | 5 days | 7/08/2024 | 7/12/2024 |
| Data Collection and Processing | 5 days | 7/15/2024 | 7/19/2024 |
| Model Development | 5 days | 7/22/2024 | 7/26/2024 |
| Model Evaluation | 3 days | 7/29/2024 | 7/31/2024 |
| Deployment | 3 days | 08/01/2024 | 08/03/2024 |

The above chart represents our timeline of our project. The project is anticipated to take around twenty-one business days, or around a month total.

## Evaluation Plan

At each stage of development, there will be a verification method or goal that will need to be achieved before the next stage of development can begin. At the first state of development, the requirements gathering stage we will need to obtain formal approval from all stakeholders, and they will need to sign-off on a document page which clearly states all the requirements of the application. The second stage of development is the data collection and processing stage which will require that management provide us a spreadsheet of data points which consist of the gender of the policyholder, the age of the policyholder, if the policyholder is a smoker or non-smoker, part of the region of the United States they reside, the number of children of the policyholder, the body-mass index of the policyholder, and total healthcare costs DFW Insurance has incurred from the policyholder. All that is required at this stage is a manual review of our data to make sure it fulfils our application requirements, and the categorical numbers can be transformed into numerical categorical. The next stage is the model development stage. To fulfil this stage of development, models will be developed, and a code review will be done to check to make sure each model works correctly. At this point a model will not be selected, that will be in the next stage. The next stage is the model evaluation where we will select the model we will be using based on the coefficient of determination (R²). Whichever model is closer to one, is the model we will select. At this stage, the application will be presented to the stakeholders, and they will need to approve it before moving on to the last stage. The last stage of development is deployment where will move the model into production and let our insurance agents evaluate the application. To complete this stage, user acceptance testing will need to be done and verified. Validation of the model will be approved by management if the model can predict a customer’s estimated healthcare costs within 25%.

## Resources and Costs

1. Labor Costs
   1. Software Engineer
   2. Hourly Rate: $60/hour
   3. Total Hours: 840 hours
   4. Total Labor Costs: $50, 400
2. Hardware Costs
   1. Laptop
      1. $2,000
3. Software and Tools
   1. Jupyter Notebook: Free (open-source)
   2. Python Libraries: Free (open-source)
4. Cloud Services (AWS)
   1. Cost per hour: $ 0.192/hour
   2. Hours used: 520
   3. Total cost: $99.84
5. Cloud Storage
   1. Cost per GB per Month: $0.023/GB
   2. Data storage Requirements: 500 GB
   3. $11.50
6. Contingency
   1. 10% of estimated budget in case of unexpected expenses: $5,251.13
7. Total Estimated Cost
   1. $57,762.47

Above is a breakdown of the estimated cost of the project, the project is estimated to cost $57,762.47 to DFW Insurance.

# Part C: Application

Files will be submitted physically along with the proposal and can be found on GitHub.

GitHub link where files are hosted: https://github.com/jaypowellcs/HealthcarePredictionModel

List of files in the application that are needed:

* 1. HealthCareCostPredictionModel.ipynb
  2. Insurance.csv

# Part D: Post-implementation Report

## Solution Summary

DFW Insurance had contacted its’ internal engineering department to try and develop a tool that can predict the healthcare costs based on a short questionnaire that their insurance agents can fill out. DFW Insurance had given us data from previous customers with a short number of attributes that they have been able to collect over the years. These data points that DFW Insurance has provided us are the age of the policyholder, the gender of the policyholder, the body mass index of the policyholder, the number of children of the policyholder, if the policyholder is a smoker or non-smoker, what part of the United States the policyholder is in based on region, and the total healthcare costs DFW Insurance has incurred by the customer. Management wanted us to access what tool would be best used to create a predictive model which takes in data that our agents can then ask and can be turned into a questionnaire. Using these data points, the software engineering team decided on a supervised learning model using a random forest regressor from scikit-learn which is an open-source machine learning library. The decisions to select this library is based on budget restrictions, since it would take a long time and resources to create our own machine learning algorithms and python was chosen due to most of our developers being familiar with the python language. We tested and trained our model against the internal data points DFW Insurance had provided us to build a prediction model where the agent can predict the total costs of healthcare costs based on the age of the policyholder, the gender of the policyholder, the body mass index of the policyholder, the number of children of the policyholder, if the policyholder is a smoker or non-smoker, and what part of the United States the policyholder is in based on region. This questionnaire can be implemented by the agent with the customer and should give the agent a relatively accurate healthcare cost prediction that the agent can in turn use to determine what price their health premium should be at. This will lead to more accurate pricing for customers, which in turn will gain DFW Insurance more customers and alleviate losses as higher premiums will be charged for higher cost customers.

## Data Summary

Provide the source of the raw data, how the data was collected, or how it was simulated. The data was provided internally from the company from previous customers. This gave us a list of one thousand, three hundred thirty-eight previous customers. The attributes that each customer had was the age of the policyholder, the gender of the policyholder, the body mass index of the policyholder, the number of children of the policyholder, if the policyholder is a smoker or nonsmoker, what part of the United States the policyholder is in based on region, and the total healthcare costs DFW Insurance has incurred by the customer. The data did have to be transformed during our application process due to it having categorical data points instead of numerical which our model would not be able to process it via one hot encoder library. Below is a link to a virtual copy of the data and a copy of an excel sheet can be found on the GitHub page and was submitted to the client.

Link to data: https://www.kaggle.com/datasets/simranjain17/insurance/data

## Machine Learning

Since we are using scikit-learn, which is an open-source machine learning library to save on time and resources instead of building our own models from scratch. I selected two models to test which was the Ridge Regression Model and a Random Forest Regressor model. I selected these after exploring the scikit-learn documentation and the machine learning roadmap that they suggested which I have included in the reference pages. Ridge regression uses a linear least square function and random forest regressor uses random tree decisions and combines them to build a prediction. After building both models and then training and testing both our models on our data, we compared both model scores by their coefficient of determination (R²) to determine which model to use. The Ridge Regression model performed worse than the random Forest Regressor model as the Ridge Regression model scored a coefficient of determination (R²) of .73, while the Forest Regressor model scored a coefficient of determination (R²) of .86. The closer to one you are, the more accurate the prediction. With the more accurate score, the Forest Regressor model was chosen.

## Validation

Since this is a supervised regression model, I judged the model by its coefficient of determination. The coefficient of determination (R²) is a number between 0 and 1 that measures how well a statistical model predicts an outcome. Our model score of a R² score of .86 means eighty six percent of the variance in the model can be explained which is a high score, which means our model has a high predictive power. However, fourteen precent of the model can not be explained by our application which could be caused by other factors or inherent randomness.

## Visualizations

Visualizations can be found in the application in part 3. And I also have included them here.

1. A screenshot of a graph

   Description automatically generated
2. A screenshot of a computer

   Description automatically generated
3. A screenshot of a graph

   Description automatically generated
4. A screenshot of a computer

   Description automatically generated
5. A screenshot of a computer screen

   Description automatically generated

## User Guide

1. **Download the required software.**
   1. To be able to run the application, you will need to download an application called Anaconda Navigator to be able to run the program. I will put a link below explaining how to download it. Please follow the steps in the link installing the program.
   2. <https://docs.anaconda.com/free/anaconda/install/windows/>
2. **After you have downloaded the software, you will need to put two files in a folder on your computer. There are two methods of obtaining the files.**
   1. GitHub – the files are hosted on GitHub, and you can download straight from the link below.
      1. Link to GitHub: <https://github.com/jaypowellcs/HealthcarePredictionModel>
   2. The files will also be submitted to the client and can be manually downloaded.
3. **The files needed are:** 
   1. HealthCareCostPredictionModel.ipynb
   2. Insurance.csv
4. **After the files are downloaded and put on a folder on your computer. Open Anaconda Navigator and click on Jupyter notebooks. I have included screenshots of what you need to select.** 
   1. After opening Anaconda Navigator, you will need to select the Jupyter Notebook option as seen in 4b.
   2. A screenshot of a computer

      Description automatically generated
   3. After selecting Jupyter Notebook you will be able to see all the files on your computer and navigate to the folder you downloaded the files into and open HealthCareCostPredictionModel.ipynb as seen in 4d.
   4. A screenshot of a computer

      Description automatically generated
5. **After the file is open, follow the instructions that are included on the top of each command, by pressing shift and enter at the same time.**
6. **To make a prediction based on the model, navigate to section 2 of HealthCareCostPredictionModel.ipynb and input the data of the customer that you would like to predict their healthcare costs. I have included an example below in 6a.** 
   1. **A screenshot of a computer

      Description automatically generated**

# Reference Page

“Choosing the Right Estimator.” *Scikit*, scikit-learn.org/stable/tutorial/machine\_learning\_map/#. Accessed 9 June 2024.