Nanyang Polytechnic

School of Information Technology

To: Dr Koh Noi Sian

Senior Lecturer

School of Information Technology

From: Yap Dong Sam (Team Leader)

210889C

Chee Bo Yu

214322H

Mark Alfonso

211684C

Tristan Hoh

212808Z

Subject: Final Project Report

Last Report

Amendment Date: 24th August 2022

# Table of Contents

[Appendix 3](#_Toc112271565)

[Executive Summary 10](#_Toc112271566)

[Modelling 15](#_Toc112271567)

[Conclusion 38](#_Toc112271568)

[Codes 40](#_Toc112271569)

# Appendix

|  |
| --- |
| **Finance (Bo Yu)** |
| Figure 1A  Figure 1B  Figure 1C    Figure 1D    Figure 1E |
| **Manpower (Tristan)** |
| Figure 2A    Figure 2B    Figure 2C Figure 2D |
| **Prevention (Mark)** |
| **[ Figure 3A]**    **[Figure 3B]**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Partitioning | | | | | | **Training** | **Validation** | **Validation**  **KS (Youden) (Logistic Regres)** | **Validation**  **KS (Youden) (Decision Tree)** | **Validation**  **KS (Youden) (SVM)** | | 50% | 50% | 0.4589 | 0.4640 | 0.3991 | | 60% | 40% | 0.4613 | 0.4658 | 0.4086 | | 70% | 30% | 0.4652 | 0.4696 | 0.4196 | | 80% | 20% | 0.4678 | 0.4791 | 0.4293 | | 90% | 10% | 0.4596 | 0.4651 | 0.4205 |   **[ Figure 3C ]**   |  |  | | --- | --- | | Support Vector Machine (SVM) - Tuning | | | **Validation KS (Youden)** | **Kernel Function** | | 0.3012 | Linear | | 0.3145 | Quadratic | | 0.4293 | Cubic |   **[ Figure 3D ]**   |  |  |  | | --- | --- | --- | | Decision Tree - Tuning | | | | **Validation KS (Youden)** | **Maximum Branches** | **Maximum Levels** | | 0.4719 | 2 (default) | 6 (default) | | 0.4726 | 3 | 6 | | 0.4791 | 3 | 8 | | 0.4786 | 3 | 10 | |
| **Accessibility (Sam)** |
| **[ Figure 4B ] [ Figure 4A ]**      **[ Figure 4C ]**      **[ Figure 4D ]**        **[ Figure 4F]**  **[ Figure 4E ]** |

# Executive Summary

**Main Objective**: “*To improve the quality of providing healthcare services.”*

|  |  |  |
| --- | --- | --- |
| **Business Objective** | **Description** | **Recommendation & Insights** |
| Finance (Bo Yu) | Profile the segment of customers that gives the highest average hospital bills (**Clustering**)  Identify factors that will affect customer cost of hospital bills  (**Linear Regression**)  **Success Criteria:** To successfully help patients to reduce their hospital bills based on the findings | Factors/Inputs that **affect** the hospital bills:   * **Age** * **Smoker** * **BMI** * **Children**   To **improve** the quality of providing healthcare services:   * **Implement Healthcare Scheme** for Aging People and Families with Children * **Depend on Patients** for Smoker and BMI   Used **Insurance(HealthCare) Dataset** for Data Modelling. Based on the Data Modelling result, I have identified the **factors** that contribute to the **higher hospital bills** and found the **segment** of customers' cost of **hospital bills.** Having this result is found, it will help patients with their **hospital bills**. Implementing a **Healthcare Scheme** will **improve the quality of healthcare services**. |
| Manpower (Tristan) | Find factors that influence the nurse to population ratio across many countries, and determine what manpower-related factors influence a country’s efficiency in handling the COVID-19 situation (namely the ratio of serious to all active cases) | Key manpower factors and their effect on the severity of cases:  Healthcare expenditure (% of country’s GDP), healthcare expenditure per capita, nurse & midwives per 1000, tests per million  Healthcare expenditure per capita also influences the manpower rate heavily  Recommendation: More money spent on healthcare guarantees that we will have more nurses. Although the objective is to have low manpower and a low number of serious cases, there is always another factor influencing how the country is handling the situation.  The Clustering Model used assesses the ratio of nurses and compares it to the serious/critical rate. We can then see an optimal ratio for how many nurses are required (in this case, the result is 12 nurses and midwives/1000 people.) We can say this is a safe ratio for optimal performance. |
| Prevention (Mark) | To determine the **factors that can reduce** hospitalization for diseases in Singapore.  **Success Criteria**: Successfully identify the factors that contribute to such diseases and recommend ways to mitigate them. | Discovered **diseases** that caused the most deaths in Singapore from Data Understanding.  Used **Cardiovascular Diseases** for Data Modelling. Discovered the top three factors that contribute to **Cardiovascular Diseases**:   * *ap\_hi* * *age(in years)* * *cholesterol*   Mitigating these factors would reduce the prevalence of Cardiovascular Diseases. Thus, resulting in a decrease in hospitalization rates. The decrease in hospitalization rates would allow hospitals to accommodate more patients with pressing illnesses and would ***improve the quality of providing healthcare services***. |
| Accessibility (Sam) | Assess the **level of inclusivity** offered, and the **overall level of Quality of Life** (QOL).  **Success Criteria**: To identify the factors that need improvement for several scopes of accessibility to support the patient's mental and physical well-being. | Factors that affect the **Accessibility of Healthcare Services** to focus on:   * Business Freedom * Economy * Housing   Ways to effectively increase the **Accessibility of Healthcare**:  **- More Business Freedom from the country** ( Increase the efficiency of government regulation of business)  **- Reduction of Economy of the country (** Decrease of all activity related to the production, consumption, trade and distribution of available goods and services)  - **Better Housing in the country** (Improvement of the Social and Physical environment where the house is located) |

# Modelling

|  |
| --- |
| **Finance (Bo Yu)** |
| **Healthcare Insurance** link: [**Insurance(HealthCare) | Kaggle**](https://www.kaggle.com/datasets/daminitiwari/insurance)   * This dataset contains the risk underwriting in Health Insurance, the interplay of **various attributes** of the insured and sees how they **affect** the insurance premium.   Modelling techniques used: **Clustering** and **Linear Regression**   * Using **Clustering modelling techniques** to profile the segment of customers that gives the highest average hospital bills   - **Changing K Clustering** to better understand the data (**From 5 Clusters to 3 Clusters**)   * Using **Linear Regression modelling techniques** to identify factors that will affect customer cost of hospital bills   - **Partitioning** was used to avoid overfitting of the model. (**70% Training, 30% Validation**)  **Tuning Process for Clustering**   * ***[ Refer to Figure 1A]*** for Changing Data Aggregation from **Sum to Average** on **Charges** * ***[Refer to Figure 1B]*** for changing K-Cluster from **5 to 3 on Clustering Model**   **Parallel Coordinates Plot**   * ***[Refer to Figure 1C]*** for inputs used for **Clustering Model** * ***[Refer to Figure 1E]*** for **5 Cluster Result**       **3 Clusters**  Based on the number of clusters to choose for my model, we can see that **3 Clusters** explain a better model for those who smoke and those who don't smoke and those who claim and don't claim insurance. By using 3 Clusters, I can **differentiate** the cluster id even better than 5 Clusters as there are too many Cluster IDs around the columns for 5 Clusters. By using a **lesser cluster** for the model, I can easily **identify simpler similarities** to interpret for my model.  The chart above shows the overview of **Cluster ID 1 to 3** where they are grouped by **age, bmi, charges, children, smoker and insuranceclaim.**    **Auto Chart**  Using Derive Cluster ID Item features, I can create the auto-chart with **Cluster ID** and **charges** as the inputs to find the highest average charges to fulfil my objectives.  Based on the charts above, **Cluster ID 3** has the highest average charges with **33291.532711.**  **Cluster ID 2 Result**    **Cluster ID 3 Result**    Based on the charts and centroids tables above, **Cluster-ID 2 customer’s** age range around **41**, bmi around **33**, belongs to the **lower range of charges**, with at least **1 child**, who is **not a smoker** but **claims insurance.**  Lastly, **Cluster-ID 3** age range around **39**, bmi around **31**, belongs to the **middle range of charges**, with at least **1 child**, who is **a smoker** and **claims insurance.**  **Tuning Process for Linear Regression**   * ***[Refer to Figure 1C]*** for Partition of **70%** Training, **30%** Validation and the random seed of **666,666**   **Fit Summary for Linear Regression**    **Before Partition**      **After Partition**      **Results**  Based on the images above, I have created a partition of **70%** training and **30%** validation and random seed of **666,666** to ensure that I can get the same training and validation data sets, to avoid overfitting of the model and to accurately evaluate my model.  Before tuning the model, the Adjusted r-square is **0.7488** and after tuning the model the adjusted r-square is **0.7487** which means **74%** of predicting the outcome. Based on the chart and Parameter Estimates above, we can analyze that if you don’t smoke it will reduce the hospital bills by **23554.3,** If you are getting older, your hospital bills will increase by **255,8591**, lastly, if your BMI is increasing, your hospital bills will increase by **329,2192.**  **Influence Plot for Linear Regression**    This **Influence Plot** shows the outliers of the case and how it affects the coefficient, the images on the right show the **top 1** influence plot, based on the result those who age 54 and have zero children are the outliers of the model and it can be further explained using the **Residual Plot.**  **Residual Plot for Linear Regression**  This **Residual Chart** shows that it is a **good model** as the data plots are closer to 0 although it is a good model, we can remove the outliers to ensure that the data plot is close together. This image above shows the location of the data plots from the top 1 influence plot and the values.  **Assessment for Linear Regression**    These line plots show model predictions versus the actual response amount in the data. Both outcomes and predictions are binned into percentiles, where it is underpredicted from the **0** and **80 to 100 Percentiles. Around 70 to 80 percentiles** of the data points are close to each other. The validation observed average is **44,453,1756** with **401 Observations Used.**    **Validating and ensuring the model is reliable**  **Centroids**  Based on the **Clustering model**, the Parallel Coordinates Plot shows that Age has a lot of inputs which means the **most important inputs**. It can also reflect in the Linear Regression Fit Summary which shows the important inputs and ensures my model is **reliable**. Those customers age around **39** and with bmi around **31**, it shows the charges are higher than the other two clusters**.**      Based on the **Linear Regression model,** we can improve the **R-Square** by adding informative missingness, using the backward variable selection method, witha **0.01** significance level to ensure my model is **reliable** and the images below show that the Adjusted R-Square has improved after adding the tuning. Having improved adjusted r-square will ensure it can **better predict** the outcome. |
| **Manpower (Tristan)** |
| Datasets used: [World Bank WDI 2.12 - Health Systems | Kaggle](https://www.kaggle.com/datasets/danevans/world-bank-wdi-212-health-systems)  [COVID-19 Dataset | Kaggle](https://www.kaggle.com/datasets/imdevskp/corona-virus-report)  Health systems dataset- contains information about expenditure, nurse count, surgical specialist count, population count for all countries  COVID-19 dataset- contains information about deaths, cases, recoveries, active cases, deaths per million etc. for all countries  Merged these datasets by country using Power Query, removed rows with a ton of nulls. Also removed unnecessary columns (e.g. alternate names for countries). These datasets are used to analyze which countries have sufficient manpower or testing ratios which let them handle the situation well.  After cleaning the dataset, it had only 108 rows, which I duplicated 1000 times for proper training. ***[See fig 2A]***  Models used  I used Linear Regression, Clustering and Decision Tree for this study.  Linear Regression: I used this to test how well the model predicts my variables, using the ratio of nurses to population as the response variable. The r-square for this model is 0.7180, hence the model is relatively reliable.  I tried using a partition over various training levels (10% to 90% in intervals of 5%), but the model actually became worse. Hence, I decided not to use one.  Fit Summary    All the variables were very relevant, as shown above. All p-values are very low  Influence Plot      Residual Plot    Very few outliers, outliers are still close to zero.  Assessment Plot    The assessment plot shows that the model will predict the values very well until the 40th percentile. Even then, the differences between the predicted and observed average is less or equal to 1, for most of the plot. After tons of rounds of testing using different variables, this was the best outcome, which is basically all the numerical variables except for raw death/cases/testing stats, as those don’t provide any real information. The death/case rate per million is a better input.  Original model can be seen in ***[Fig 2B]***, using only the variables I thought were most relevant (as derived from the Correlation Matrix). The model was very unreliable with a low r-square and bad assessment plot.  Clustering: I used this model to derive the best values of variables across the clusters. Using the parallel coordinates plot, I am able to see the relationship of the different variables, as well as how many observations go into each cluster.  Cluster Diagram    ***[See Fig 2C].*** Variables used: Serious/critical ratio, nurse/midwife ratio, population.  I kept it to these variables as using more would result in too many polylines. As for why these were chosen, I wanted to pick variables that would emulate a certain demographic, so we can see the average stats for a cluster based on the country’s circumstances- the population can influence the spread of the virus, thus by comparing clusters with similar population I could extract other variables for a fairer comparison.  Before and After of Parallel Coordinates Plot    This plot shows the relationship between the serious/critical ratio and the nurse/midwife ratio. It is mostly inversely proportional, and most of the cases go into Cluster 4 and 6 which generally have a low serious/critical ratio.  Default: 5 clusters, each cluster was imbalanced and the observations did not provide very useful insights. Tried less and more clusters, all resulted in imbalanced or insufficient data for each cluster. Hence, I used 6 as it provided the best comparison. The parameters were set as shown- ***[Fig 2C]***    By isolating the cluster plot for serious/critical ratio and nurse to population ratio, we can analyze the clusters that have the best ratios, and determine an optimal ratio for a country to handle the situation best. According to the graph, cluster 6 features the most in the bottom left of the graph (which means low manpower and low serious cases), hence we can use this cluster and check other statistics for it.  Another insight we can gather is on how many nurses are needed to keep the number of serious cases low- in this case, there are high serious/critical ratios for the range of 0-10 nurses per 1000 people. This means a safe estimate is roughly 12 nurses/1000 people.  For instance, the Expenditure statistic for cluster 6 is way higher than average, which could explain its efficiency in handling the situation. Conversely, the ratio of surgical specialists to 1000 people isn’t as high for this cluster. Hence, it is possible that this variable is irrelevant for this case.  Autochart plots for Healthcare Expenditure GDP and Surgical Specialists      Clusters 4 and 6 have similar serious/critical ratios and population, hence we can compare them.  Decision Tree: I used this to find the most significant variables in relation to the nurse to population nurses    The most significant variable here is health expenditure per capita. I only selected a few variables for this model for a clearer comparison, and the other variables were largely insignificant. ***[See Fig 2D for setup]*** |
| **Prevention (Mark)** |
| Cardiovascular Diseases dataset: <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>  The dataset contained 70,000 records and was generally balanced. The response feature, ***cardio***, contained around 35,000 records for both **1** and **0** inputs.  With reference with ***[Figure 3A]***, All features were useable for Data Modelling:   * **Numeric Features:** *age (in years), height, weight, ap\_hi, ap\_lo* * **Categorical Features:** *gender, cholesterol, gluc, smoke, alco, active, cardio*   Changed the type of the features accordingly in SAS Viya.  With reference with ***[Figure 3B]***,I paritioned the dataset into **80% Training and 20% Validation** as the validation KS (Youden) of the models are the best. I set the seed as **11223344** to ensure that the model results stays the same every reload.  **Model Overview**  Modeling techniques used: **Logistic Regression, Support Vector Machine (SVM), Decision Tree**     * The Model Overview above shows the model comparison of the three models. The best model is the Decision Tree model as it has the highest **Validation KS (Youden), 0.4791**. * The worst model is SVM as the **Validation KS (Youden), 0.4293,** is the worst. This is already with the **Cubic** Kernel power, the highest kernel power SAS Viya can provide. (refer to ***[Figure 3C]***) * Since the Validation KS (Youden) of SVM is not on par with the other two models, I will exclude it from the model explanation.   **Decision Tree**  There was not much tuning done in the decision tree. I only changed the branches and the levels of the tree. (refer to ***[Figure 3D]***)   * I changed the maximum branches to 3 as there are features with 3 categories. This would expand the decision tree more and allow us to better analyse the features. * I also changed the maximum levels to 8 as its has the highest Validation KS (Youden) amongst the other configurations.      * Now we have the decision tree, I have to identify the leaf nodes that has at least 70% of its response = 1, and have at least 100 counts in that node. This increases the reliability in determining the relevant factors based on the split rules. * We can identify the leaf nodes using the **leaf statistics**.      * From the leaf statistics, we can observe that there are 8 nodes that have at least 70% of their response = 1. We must now filter the nodes that have less than a hundred counts.      * Now we have identified the relevant leaf nodes, we can conduct further analysis on the remaining leaf nodes. We can observe that the relevant features that are commonly seen are: *age(in years), ap\_hi, cholesterol*. * These factors can further be supported by the Variable Importance chart.      * We can observe that the top 3 features shown in the Variable Importance chart supports the features discovered from the leaf node analysis. We can support the features further using the Logistic Regression’s Fit Summary.      * The top three factors ranked by variable importance are also *ap\_hi, age(in years), cholesterol*.   **Model Comparison**     * When comparing the confusion matrix, the accuracy and precision of the Decision Tree model is 73.93% and 76.47% (to 2dp) respectively. * The accuracy and precision of the Logistic Regression model is 72.93% and 76.32% (to 2dp) respectively. * Decision Tree and Logistic Regression models are both relatively close in accuracy and precision.      * From the Lift Chart, we can observe that the Decision tree model is the best as the model is more effective at the 30th percentile, even though logistic regression is the better model at the 5th percentile.      * From the ROC Chart, the Decision Tree model is the best model as it is closer to the true positive rate. This is followed by the Logistic Regression model. * This tells me that the Logistic Regression model is acting like a support to the Decision Tree model, which increases the accuracy of the results, making them more reliable. |
| **Accessibility (Sam)** |
| Quality of Life Link: <https://www.kaggle.com/datasets/orhankaramancode/city-quality-of-life-dataset>  The dataset was mostly clean and balanced.  Response Variable: **Healthcare** (UA Score for Healthcare ranging from 1 to 10)  Input Variables for Decision Tree: [ refer to **Figure 4A** ]  Input variables for Linear Regression: [ refer to **Figure 4B** ]  Modeling techniques used: **Decision Tree** and **Linear Regression**   * Using **Decision Tree** to identify the best factors that influence the accessibility of healthcare. * Using **Linear Regression modelling techniques** to find the intensity of different factors that affect the accessibility of healthcare. * Partition the dataset into **70% Training and 30% Validation**. Used Partitioning for all Models. * Some numeric variables from the “Continuous effects” of the Linear regression model are removed because they are statistically insignificant.   Parameters Tuning for **Decision Tree**:  The validation average square error had reduced by 0.1264 with the following Parameters Tune:  - Maximum Branches increased from 2 to 3  - Maximum Levels of the Decision Tree were increased from 6 to 8  - Predictor Bins increased from 50 to 300.  (Having a minimal average square error in the model is pivotal in ensuring the best model is used for the Decision Tree) [ refer to **Figure 4C**  and **Figure 4D** ]  Validation process of **Decision Tree**:   * Usage of the Partition ID to split into 70 % Training Dataset and 30% Testing Dataset for a more accurate result.   Modelling Process of **Decision Tree**:  The Decision Tree has its distributed leaf nodes as shown with the relevant parameters set.  Leaf nodes 106 and 128 have the highest average healthcare score of 9.09 and 9.06 respectively.    The relevant leaf nodes have the respective decisions made above.      The Assessment shows the observations used for the Training and Validation Partition. The predicted and the observed average seems more fluctuated between the 40 and 80 percentile  Parameters Tuning for Linear Regression:  Shows more variables that are not significant for use in the Linear Regression Model (Ensuring the best model in using the most significant variables for analysis) [ refer to **Figure 4E**  and **Figure 4F** ]  Validation Process for Linear Regression:   * Usage of the Partition ID to split into 70 % Training Dataset and 30% Testing Dataset for a more accurate result.      * The Input Variables are able to explain 72.46 % of the Healthcare Quality of Life Score.   Modelling Process for Linear Regression:    All statistically significant variables are used for the linear regression model (p-value < 0.05). All variables that have multicollinearity are removed from the model.    The parameter estimates show the estimate between the “Parameter” and the Healthcare variable in descending order. As shown above, there is a **Direct Relationship** between UA\_Continent Oceania (Categorical Variable) and Healthcare. For every **1 unit** of increase in **UA\_Continent Oceania**, **Healthcare** would have an increase of **4.31 units** (Vice Versa).  Moreover, there is a **Direct Relationship** between Business Freedom and Healthcare. For every **1 unit** of increase in **Business Freedom**, **Healthcare** would have an increase of **4.12 units** (Vice Versa)    Most of the points are situated together. Minimal Outliers are situated in the Plot which makes Linear Regression a good model.    The Assessment shows the observations used for the Training and Validation Partition. Compared to the Decision Tree, the linear regression has more fluctuating results for predicted and observed average |

# Conclusion

***To improve the quality of providing healthcare services***, we need to look at the conclusions of the four business objectives:

|  |  |
| --- | --- |
| **Business Objectives** | **Conclusion** |
| Finance (Bo Yu) | By finding the factors that contribute to **higher hospital bills** and identifying the segment of customers that gives the **highest average hospital bills**, patients can **better understand** the cost of their hospital bills and healthcare can **help** patients with support for those **in need** by **implementing a healthcare scheme**. This will eventually give healthcare a **better quality** of providing healthcare services as the healthcare will know who to focus more on. |
| Manpower (Tristan) | By exploring the factors that influence the **Nurse to Population Count** and the ratio of **Serious and Critical to Active Cases**, we can optimize these ratios to maximize staff efficiency.  Identifying the group of countries that is currently handling the situation best can help us create a system that ticks all the requirements for efficiency and other factors, and we can do this by taking a look at that group’s statistics, such as population and expenditure, and use those to draw a conclusion based on good estimates for all those parameters. All in all, it helps to prevent understaffing/overstaffing while keeping the amount of serious cases low. |
| Prevention (Mark) | By mitigating factors that contribute to Cardiovascular Diseases, we can reduce the prevalence of those diseases. The decrease in prevalence would result in a decrease in hospitalisation rates. This would lead to an increase in the accommodation of patients in hospitals, which will ultimately improve the quality of providing healthcare services. |
| Accessibility (Sam) | By knowing the relevant factors to improve the accessibility of healthcare, these can further support the patient's mental and physical well-being. The result obtained proves that better business freedom, reduction in the economy, and better housing increase the accessibility of healthcare. With all these considerations of a country, this improves the level of inclusivity offered, and the overall level of Quality of Life (QOL) for patients in Healthcare. |

The conclusions of the Business Objectives are all **positive** and do contribute to answering the business problem.

# Codes

|  |  |
| --- | --- |
| **Business Objectives** | **Codes** |
| Finance (Bo Yu) | Note: Based on the dataset found for Finance, there wasn’t much tuning required except for changing the data type for some of the inputs used like changing from **Sum to Average for insuranceclaim, sex and smoker columns**. Lastly, changing from **Measure to Category** for the inputs in SAS VIYA and **adding new columns** called **insuranceclaim** in excel. |
| Manpower (Tristan) | For the dataset I used, only basic cleaning was required, and I only had to use Power Query for merging.  All datatypes were changed in Excel and SAS, and I did manual imputation for bugged values as there were only a few. |
| Prevention (Mark) | Note: The Cardiovascular Diseases Dataset did not require much transformation and cleaning. The only thing I added to the dataset was another column, age (in years). All I did was performed a simple excel calculation with the age (in days), the given age column, and changed it to years by dividing it by 365. |
| Accessibility (Sam) | Note: The Quality of Life Dataset for Accessibility requires minimal data preparation. The only manipulation needed for the dataset was to multiple the columns by 4 for modelling to ensure data granularity. |