**BUSINESS UNDERSTANDING**

The business understanding of the malaria commodities prediction model was a crucial component of the model development process. It involved identifying the client or user, their needs, and the objectives of the task.

The client for this model was the Ministry of Health whom are responsible for procuring and distributing malaria commodities such as mosquito nets, anti-malarial drugs, and diagnostic kits. Their need was to ensure that they had the appropriate quantities of these commodities available at the right time and in the right locations to effectively respond to malaria outbreaks.

The objective of the task was to develop a machine learning model that accurately predict the demand for malaria commodities in different regions and at different times of the year. This allowed the client to better anticipate demand and plan for procurement and distribution, accordingly, ultimately leading to more effective malaria prevention and treatment.

The client engagement process for this task involved working closely with the health ministry to gather data on malaria incidence and commodity distribution, as well as understanding their specific needs and challenges. The model development process involved creating an appropriate data set (provided by MoH), selecting and training a machine learning algorithm, and fine-tuning the model for optimal performance.

The final model needd to be deployed in a user-friendly and accessible way, with the results presented in a way that is easily understandable and actionable by the Ministry of Health. Regular monitoring and evaluation of the model's performance would also be important to ensure that it continues to meet the client's needs over time.

**DATA ACQUISITION**

The data acquisition process was an important step in developing a malaria commodity demand prediction model. It involved identifying and collecting the data needed to train and test the machine learning algorithm.

The source systems for this model includes the health information system (HIS) or other databases maintained by the health ministry. These systems contain information on the distribution of malaria commodities, as well as data on malaria incidence and other relevant factors such as climate and population density.

The data acquisition process involve the following steps:

1. Identify the relevant data sources and determine how to access them.
2. Determine what data is needed for the model and extract it from the source systems. This involve working with IT staff or other stakeholders to obtain access to the data and ensure that it is in a usable format.
3. Clean the data and perform any necessary pre-processing steps. This includes removing duplicates or invalid records, filling in missing data, or transforming the data into a format suitable for machine learning algorithms.
4. Combine the data from the different sources into a single data set for analysis.
5. Split the data set into training and testing subsets, to be used to train and evaluate the machine learning algorithm.

In addition to the above steps, it was necessary to conduct exploratory data analysis to better understand the relationships between different variables and determine which factors are most important for predicting malaria commodity demand. We used yearly sum of the data for the model.

Overall, the data acquisition process is a complex and time-consuming task and required close collaboration with stakeholders and technical experts to ensure that the data used in the model is accurate, reliable, and appropriate for the intended use case.

**EXPLORATORY DATA ANALYSIS**

Exploratory Data Analysis (EDA) is an essential step in any data science project, including the development of a malaria commodity demand prediction model. The EDA process helps to better understand the data, identify patterns and trends, and inform the selection of features and algorithms for the machine learning model.

The EDA process for this model involved the following steps:

1. Understanding the data: This involved gathering background knowledge about the data, including the source of the data, the meaning of each variable, and any known issues or limitations of the data. The source of the data was MoH.
2. Checking for missing values: This involved checking for any missing data in the dataset and deciding on the appropriate way to handle them. Missing values can be imputed or removed depending on the nature of the data. Decided to fill a zero in missing values.
3. Descriptive statistics: This step involved computing summary statistics such as mean, median and total for each variable. This provides an overall understanding of the data and can help identify any outliers, skewness, or other anomalies in the data.
4. Data visualization: This step involved creating visualization such as bar charts and stacked bar charts to identify patterns, trends, and correlations between the different variables.
5. Feature engineering: This involved selecting the most important features or variables to be used in the machine learning model. This involved combining or transforming variables, as well as creating new features that better capture the underlying relationships in the data. Transformed the counties to one string of lower case variables each.
6. Correlation analysis: This step involved calculating the correlation between different variables to identify any strong relationships or dependencies. This information was used to inform the selection of features and algorithms for the machine learning model. Since it was a prediction model we decided to use linear regression algorithm.

The EDA process for the malaria commodity demand prediction model helped to provide valuable insights into the data, allowing for better feature selection and more accurate predictions. The outcomes of the EDA process, including summary statistics and data visualizations, can be presented to stakeholders to inform the development of the machine learning model.

**DATA CLEANING**

Data cleaning is an important step in the data preparation process for building a malaria commodity demand prediction model. The aim of this step was to ensure that the dataset is free of errors, inconsistencies, and irrelevant data that can negatively affect the accuracy of the model.

The data cleaning process for the malaria commodity demand prediction model may involve the following steps:

1. Identifying missing or invalid values: In this step, we identified any missing or invalid values in the dataset. This involved filling in missing values (with zeros) which we largely used or removing them altogether if they cannot be filled.
2. Handling outliers: Outliers distorted the statistical analysis and negatively impacted the accuracy of the model. We handled outliers by either removing them or transforming them using methods such as robust statistics which involved use of median instead of mean.
3. Handling duplicates: Duplicates can cause issues during the training and testing phase. We identify and removed duplicates using ‘drop\_duplicate()’ function.
4. Data normalization and standardization: We applied normalization and standardization techniques to ensure that the data is scaled and consistent across all variables.
5. Handling irrelevant or redundant features: We removed irrelevant or redundant features from the dataset so as to improve the model's performance and simplify the data processing pipeline.
6. Data encoding: In this step, we encoded categorical data into numerical data to the chatbot (our user interface). This is necessary because most machine learning algorithms can only work with numerical data.

The outcomes of the data cleaning process for the malaria commodity demand prediction model included a cleaned dataset that is free from errors and inconsistencies. The was suitable for training the machine learning algorithm, improving the accuracy of the model's predictions. Additionally, any issues or limitations discovered during the cleaning process was documented to ensure transparency and reproducibility of the data preparation process.

**FEATURE ENGINEERING**

Feature engineering is the process of selecting, creating, and transforming variables in a dataset to improve the performance of a machine learning model. The feature engineering process for the malaria commodity demand prediction model involved the following steps:

1. Feature selection: This involved selecting the most relevant features to use in the model. Relevant features are those that have a strong correlation with the target variable and can help explain the patterns in the data. For us the most relevant features were the counties, years and the data of diagnostic malaria kits.
2. Feature transformation: This involved transforming variables to make them more informative or easier to work with. Common transformations that we used included log transformations, scaling, and standardization.
3. Feature creation: This involved creating new features from existing variables. For example, we created a new variable name for the counties that improved accessibility of the diagnostic malaria kits data.
4. Feature extraction: This involved extracting new features from the existing variables. For example, we extracted the trends over the years in the data using linear regression analysis techniques.
5. Feature scaling: This involved scaling the features to ensure that they are all on a similar scale by fitting the features. This improved the performance of linear regression algorithms.

The features for the malaria commodity demand prediction model include variables such as the counties the kits were distributed to and the period in months and years of diagnostic malaria kits dispensed by the MoH. Other potentially relevant variables could include the weather conditions, socio-economic indicators, and demographic information which were not available in the data set.

The feature engineering process helped to create a more informative and effective set of features for the machine learning model. By selecting and transforming the most relevant features, we improved the accuracy of the model's predictions and provide more valuable insights into the factors that influence malaria commodity demand.

**MODEL DEVELOPMENT**

Model development is the process of selecting an appropriate machine learning algorithm and tuning its hyperparameters to optimize its performance on the given dataset. For the malaria commodity demand prediction model, we can use a range of machine learning algorithms, such as linear regression, decision trees, random forests, and neural networks. The choice of the algorithm and the hyperparameters depends on the nature of the data, the complexity of the problem, and the desired performance of the model.

The model development approach for the malaria commodity demand prediction model may involve the following steps:

1. Splitting the data: We can split the dataset into training, validation, and test sets. The training set is used to train the model, the validation set is used to tune the hyperparameters, and the test set is used to evaluate the final performance of the model.
2. Selecting an algorithm: We can try out different algorithms and evaluate their performance on the training set using metrics such as mean squared error, root mean squared error, or R-squared. Based on the performance, we can select the most appropriate algorithm.
3. Tuning hyperparameters: We can adjust the hyperparameters of the chosen algorithm to optimize its performance on the validation set. This can be done using techniques such as grid search, random search, or Bayesian optimization.
4. Model training: After selecting the algorithm and hyperparameters, we can train the model on the entire training set.
5. Model evaluation: We can evaluate the final performance of the model on the test set, using metrics such as mean squared error or R-squared.

The choice of the machine learning algorithm and hyperparameters would depend on the nature of the data and the desired performance of the model. For the malaria commodity demand prediction model, we could use a range of algorithms, such as linear regression, decision trees, or random forests, depending on the complexity of the problem and the nature of the data. The final choice of the algorithm would be based on the performance on the training and validation sets.

The justification of the chosen model would be based on its performance on the validation and test sets, as well as its interpretability and ability to provide actionable insights into the factors that influence malaria commodity demand. Additionally, the choice of the model would be based on its ability to generalize well to new data and its ability to handle missing or noisy data.

**MODEL EVALUATION**

Model evaluation is the process of assessing the performance of the machine learning model on the test set, using appropriate evaluation metrics. The choice of evaluation metrics depends on the nature of the problem and the desired performance of the model. For the malaria commodity demand prediction model, we used metrics such as mean, median and sum.

In model evaluation, there are various metrics used to determine how well the model performs. Mean, median and sum are three of the commonly used metrics for evaluating the performance of linear regression models.

The mean is the average value of a dataset. It is obtained by summing all the values in the dataset and dividing by the number of observations. In the context of model evaluation, the mean can be used to calculate the average error of the model. This metric is commonly referred to as the mean squared error (MSE), and it is a measure of how well the model fits the data. The lower the MSE, the better the model.

The median, on the other hand, is the middle value of a dataset. It is the value that divides the dataset into two equal halves. In the context of model evaluation, the median can be used to determine the central tendency of the errors. This metric is commonly referred to as the median absolute error (MAE), and it is a measure of how well the model fits the data. The lower the MAE, the better the model.

The sum, as the name implies, is simply the total value of a dataset. In the context of model evaluation, the sum can be used to determine the total error of the model. This metric is commonly referred to as the residual sum of squares (RSS), and it is a measure of how well the model fits the data. The lower the RSS, the better the model.

In summary, the mean, median, and sum are three commonly used metrics for evaluating the performance of linear regression models. The mean is used to calculate the average error of the model, the median is used to determine the central tendency of the errors, and the sum is used to determine the total error of the model. The choice of which metric to use depends on the specific use case and the requirements of the model.

**MODEL DEPLOYMENT**

Model deployment is the process of making the machine learning model available for use in a production environment. The choice of deployment platform depends on various factors such as the application requirements, the scalability and performance requirements, and the budget.

In the case of the malaria commodity demand prediction model, we deployed the model on VPS (Virtual Private Server). VPS platforms offers a range of services for deploying, scaling, and managing machine learning models, as well as tools for monitoring, debugging and hosting.

The process of deployment involves several steps, including packaging the model and its dependencies, creating a Flask restful API for the model, and deploying the model to the chosen deployment platform which is VPS.

Once the model is deployed, it could be monitored and updated regularly to ensure that it continues to perform well in the production environment. Monitoring tools can be used to track the performance of the model and detect any issues or anomalies. The model could also be updated periodically to incorporate new data and retrain the model to improve its performance.

Overall, the process of model deployment should be well-planned and documented to ensure that the deployed model is reliable, scalable, and easy to maintain.

**CHALLENGES**

There are several challenges that we faced when creating the model to predict demand for malaria commodities, including:

1. Data quality: One of the biggest challenges in building the model was ensuring that the data used to train the model is accurate, complete, and representative of the underlying phenomenon. In the case of malaria commodities, the data was partly incomplete, contained errors or outliers and was subject to measurement or reporting bias.
2. Seasonality and trends: Another challenge was dealing with seasonality and trends in the data. Demand for malaria commodities varied over time due to factors such as weather, disease outbreaks, and public health campaigns. Accounting for these factors was difficult, especially if there were multiple interacting factors that influence demand.
3. Confounding variables: There were other factors that influenced demand for malaria commodities that were not captured by the available data, such as changes in population, access to healthcare, or socio-economic status. These factors introduced noise and bias into the model, making it less accurate.
4. Model complexity: The choice of model used to predict demand also posed challenges. A model that is too simple may not capture all of the relevant factors that influence demand, while a model that is too complex may be overfitting to the noise in the data, leading to poor performance on new data.
5. Implementation challenges: Finally, deploying the model in a real-world context posed its own set of challenges. For example, the model need to be integrated with other systems, such as supply chain management or inventory tracking software, and the output of the model need to be communicated effectively to stakeholders who make decisions about restocking and distribution.