Logistic Regression Analysis for Heart Health Data: Predicting Delay in Seeking Medical Treatment

Issues:

1. Overfitting: Model is too complex, it may fit the training data too well and perform poorly on new, unseen data. Therefore, it is crucial to avoid overfitting by using appropriate regularisation techniques.
2. Confounding variables: The selected predictor variables may be correlated with other variables not included in the model, which could result in spurious associations and inaccurate predictions.
3. Missing data: If missing values are in the dataset, it may be necessary to blame them to avoid bias in the model.
4. Model interpretation: It is essential to carefully interpret the coefficients and statistical significance of the predictor variables in the logistic regression model. In particular, categorical variables may require special attention regarding their interpretation and encoding.
5. Generalizability: The logistic regression model may generalise poorly to other populations or settings outside the data used to train the model. Therefore, validating the model on independent datasets is vital to ensure its generalizability.

Findings:

1. A logistic regression model was built to predict whether a person seeks medical treatment in 2 days or less ("1") or takes longer than two days to seek medical treatment ("0").
2. The median number of delay days in seeking medical treatment was 2.
3. The predictor variables in the heart health dataset include 17 categorical variables and one continuous variable (age).
4. The most useful predictor variables in the logistic regression model for predicting the outcome may depend on the specific analysis and scenario considered.
5. Different logistic regression models may be built to predict whether a person seeks medical treatment on or less than the cohort average delay days or on

or less than one day, and the most useful predictor variables may also differ in these scenarios.

Discussions:

1. Interpretation of predictor variables: Depending on the specific analysis, some predictor variables may be more strongly associated with seeking medical treatment within a particular time frame than others. Discussing the potential underlying mechanisms for these associations and the implications for clinical practice or future research may be necessary.
2. Comparison of scenarios: Comparing the logistic regression models for predicting different techniques (e.g., seeking treatment within two days vs. the average delay time vs. one day) could provide insights into the relative importance of other predictor variables and the trade-offs involved in different decision-making scenarios. This discussion could also include potential limitations or biases associated with each design.
3. Model performance and validation: It is essential to discuss the overall performance of the logistic regression models and their ability to predict outcomes accurately. It might include metrics such as the area under the receiver operating characteristic curve (AUC-ROC) or the accuracy and potential sources of bias or variability. Validating the model on independent datasets or conducting sensitivity analyses could also be discussed.
4. Clinical implications: The logistic regression models and findings could improve the timeliness and quality of medical care for patients with heart health issues. This discussion could include potential interventions or strategies to reduce delay in seeking medical treatment and the potential benefits and challenges associated with implementing these interventions.
5. Future research directions: The logistic regression analysis could generate new hypotheses or research questions related to the predictors of timely medical treatment seeking and potential areas for intervention. Possible directions for future research include exploring additional predictor variables, conducting longitudinal studies, or testing interventions to reduce delay in seeking medical treatment.

Appendix A- Method:

To construct a logistic model that predicts whether an individual seeks medical treatment within two days, the following steps should be taken:

1. Load the heart health dataset provided in the .xls file.

2. Create a binary outcome variable to represent individuals who seek medical treatment within two days ("1") and those who take longer than two days to seek treatment ("0").
3. Utilise the predictor variables in the dataset to develop a logistic regression model that predicts the binary outcome variable.
4. Determine the most significant predictor variables in the model.
5. Repeat the previous steps while using different cutoff points for delay days (such as the average delay days of the cohort or one day), and compare the results to assess if the most important predictor variables change.

Appendix B- Results:

The median number of delay days is 2. Build a logistic model to predict whether a person seeks medical treatment in 2 days or less (“1”) or takes longer than 2 days to seek medical treatment (“0”). Which factors does your model suggest are most useful in predicting the outcome? The result for this criteria is Accuracy: 0.5619834710743802

How would your logistic model differ if it were to predict whether a person seeks medical treatment on or less than the cohort average delay days (“1”), or takes longer than the average number of days to seek medical treatment (“0”)? Which factors does your model suggest are most useful in predicting the outcome? The result for this criteria is that the current function value is 0.583225 and the iterations are 5.

How would your logistic model differ if it were to predict whether a person seeks medical treatment on or less than 1 day (“1”) or takes longer than 1 day to seek medical treatment (“0”)? Which factors does your model suggest are most useful in predicting the outcome? The result for this criteria is the accuracy is 0.72. That means the logistic regression model is accurate 72%.

Appendix C- Code:

To Build a logistic model to predict whether a person seeks medical treatment in 2 days or less (“1”) or takes longer than 2 days to seek medical treatment (“0”),

```
import pandas as pd
import numpy as np
```

```
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
heart_data = pd.read_excel("heart-health-data.xls")
heart_data

heart_data.describe()
heart_data.corr()

#Creating dummy data
heart_data_dummy = pd.get_dummies(heart_data)
heart_data_dummy.dropna(inplace=True)
heart_data_dummy
heart_data_dummy["delaydays_2or_less"] =
np.where(heart_data_dummy["delaydays"] <= 2, 1, 0)

X = heart_data_dummy.drop(["ID", "delaydays", "delaydays_2or_less"],
axis=1)
y = heart_data_dummy["delaydays_2or_less"]
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=40)
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(y_pred)
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)
X = sm.add_constant(X)
logistic_model = sm.Logit(y_train, X_train).fit()
print(logistic_model.summary())
# Print the coefficients in descending order
coef = pd.DataFrame({'coef': model.coef_[0], 'variable':
X_train.columns})
coef = coef.sort_values(by='coef', ascending=False)
print(coef)
```

To modify the logistic regression model to predict whether a person seeks medical treatment on or less than the cohort average delay

days (“1”) or takes longer than the average number of days to seek medical treatment (“0”),

```
#First, we can calculate the cohort average delay days
avg_delay_days = heart_data_dummy['delaydays'].mean()
print(avg_delay_days)
#creating a binary outcome variable
heart_data_dummy['delaydays_outcome'] = (heart_data_dummy['delaydays']
<= avg_delay_days).astype(int)
heart_data_dummy['delaydays_outcome']

X = heart_data_dummy[['Age', 'Gender', 'Ethnicity', 'Marital',
'Livewith', 'Education', 'palpitations', 'orthopnea', 'chestpain',
'nausea', 'cough', 'fatigue', 'dyspnea', 'edema', 'PND', 'tightshoes',
'weightgain', 'DOE']]
y = heart_data_dummy['delaydays_outcome']

X = sm.add_constant(X)
log_reg_avg = sm.Logit(y, X).fit()
log_reg_avg.summary()
```

To build a logistic regression model to predict whether a person seeks medical treatment on or less than 1 day (“1”) or takes longer than 1 day to seek medical treatment (“0”):

```
heart_data_dummy['early_treatment'] = (heart_data_dummy['delaydays'] <=
1).astype(int)

# Create a list of predictor variables
predictors = ['Age', 'Gender', 'Ethnicity', 'Marital', 'Livewith',
'Education',
              'palpitations', 'orthopnea', 'chestpain', 'nausea',
'cough', 'fatigue',
              'dyspnea', 'edema', 'PND', 'tightshoes', 'weightgain',
'DOE']

subset_data = heart_data_dummy[predictors]

# Fit the logistic regression model
```

```
logit_model = sm.Logit(heart_data_dummy['early_treatment'],
heart_data_dummy[predictors]).fit()

# Print the summary table of the model
print(logit_model.summary())
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)

# Train the logistic model on the training set
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

# Predict the outcome using the testing set
y_pred = logreg.predict(X_test)

# Evaluate the performance of the model
from sklearn.metrics import confusion_matrix, classification_report
print(confusion_matrix(y_test, y_pred))

print(classification_report(y_test, y_pred))
```