```
#
# Purdue University Global
#
# IN402 - Modeling and Predictive Analysis
#
# Unit 10 Assignment / Module 6 Part 3 Competency Assessment
#
# Classification Model Selection
#
# Jupyter Notebook Code
#

# [1] ###########################################################################

# Import packages

import sys

# Ignoring warnings

if not sys.warnoptions:
    import warnings

warnings.simplefilter("ignore")

# [2] ###########################################################################

# Import the dataset

# Importing the dataset to a pandas DataFrame

import pandas as pd

df = pd.read_csv('/home/codio/workspace/data/IN402/Churn_Modelling.csv')

print( df.shape)

# [3] ###########################################################################

# Wrangle the data

# Drop columns with no analytical value

df = df.drop(['RowNumber', 'CustomerId', 'Surname'], axis=1)

# [4] ###########################################################################

# Convert the categorical columns into dummy columns
# and drop the original categorical columns

geography = pd.get_dummies(df.Geography).iloc[:,1:]

gender = pd.get_dummies(df.Gender).iloc[:,1:]

# Drop columns with non-numeric data

df = df.drop(['Geography', 'Gender'], axis=1)

# [5] ###########################################################################

# Join the dummy columns into the main dataset

# Add columns with converted dummy values

df = pd.concat([df,geography,gender], axis= 1)

# [6] ###########################################################################

# Split the dataset into target and feature subsets.

X = df.drop(['Exited'], axis=1)

y = df.loc[,'Exited']

# [7] ###########################################################################

# Select features

# Check the variance in the numeric variables

from statistics import variance

creditScore = df['CreditScore']

age = df['Age']

tenure = df['Tenure']

balance = df['Balance']

estimatedSalary = df['EstimatedSalary']

# [8] ###########################################################################

# Display the parameter variances

print("Variance of CreditScore is % s " %(variance(creditScore)))

print("Variance of Age is % s " %(variance(age)))

print("Variance of Tenure is % s " %(variance(tenure)))

print("Variance of Balance is % s " %(variance(balance)))

print("Variance of EstimatedSalary is % s " %(variance(estimatedSalary)))

# [9] ###########################################################################

# Split the dataset into training and testing subsets.

from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X, y, test_size=0.2, random_state=0)

# [10] ##########################################################################

# Display size of training set

print(len(X_train))

# [11] ##########################################################################

# Display size of test set

print(len(X_test))

# [12] ##########################################################################

# Conduct feature scaling (required by SVM)

from sklearn.preprocessing import StandardScaler

# feature scaling is required by SVC

sc = StandardScaler()

X_train = sc.fit_transform(X_train)

X_test = sc.transform(X_test)

#X_train[:,-2:] = sc.fit_transform(X_train[:,-2:])

#X_test[:,-2:] = sc.transform(X_test[:,-2:])

X_train

# [13] ##########################################################################

# Model using Logistic Regression

# Import packages

from sklearn.linear_model import LogisticRegression

from sklearn import metrics

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# [14] ##########################################################################

# Build the model

lr_model = LogisticRegression()

# Fit the model

result = lr_model.fit(X_train, y_train)

# Predict using the model

prediction_test = lr_model.predict(X_test)

# [15] ##########################################################################

# Evaluate the model

# Print the prediction accuracy

print(metrics.accuracy_score(y_test, prediction_test))

# [16] ##########################################################################

# Display the confusion matrix

print(confusion_matrix(y_test, prediction_test))

# [17] ##########################################################################

# Display the classification report

print(classification_report(y_test, prediction_test))

# [18] ##########################################################################

# To get the weights of all the variables

weights = pd.Series(lr_model.coef_[0], index=X.columns.values)

weights.sort_values(ascending = False)

# [19] ##########################################################################

# Model using SVM

# Import packages

from sklearn.svm import SVC

from sklearn import metrics

# [20] ##########################################################################

# Build the model

svm_model = SVC(kernel = "linear")

# Fit the model

# Train the model

svm_model.fit(X_train, y_train)

# Predict using the model

svm_prediction = svm_model.predict(X_test)

# [21] ##########################################################################

# Evaluate the model

print("accuracy: ", metrics.accuracy_score(y_test, y_pred = svm_prediction))

# [22] ##########################################################################

# Precision score

print("precision: ", metrics.precision_score(y_test, y_pred = svm_prediction))

# [23] ##########################################################################

# Recall score

print("recall", metrics.recall_score(y_test, y_pred = svm_prediction))

# [24] ##########################################################################

# Display classification report

print(metrics.classification_report(y_test, y_pred = svm_prediction))

# [25] ##########################################################################

# Model using RandomForestClassifier

# Import packages

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, accuracy_score

# [26] ##########################################################################

# Build the model

rf_model = RandomForestClassifier(n_estimators=200, random_state=0)

# Fit the model

rf_model.fit(X_train, y_train)

# Predict using the model

rf_predictions = rf_model.predict(X_test)

# [27] ##########################################################################

# Evaluate the model

print(classification_report(y_test,rf_predictions))

# [28] ##########################################################################

# Display the accuracy score

print(accuracy_score(y_test, rf_predictions ))

# [29] ##########################################################################

# Display the accuracy score

import matplotlib.pyplot as plt

%matplotlib inline

feat_importances = pd.Series(rf_model.feature_importances_, index=X.columns)

feat_importances.nlargest(10).plot(title="Accuracy Score", kind='barh')

# ###############################################################################
```