# D209 Data Mining 1 - NVM2 - Classification Analysis

July 24, 2021

**NVM2 Task 1: Classification Analysis**

**D209 Data Mining Performance Assessment** Ryan L. Buchanan
Student ID: 001826691
Masters Data Analytics (12/01/2020)
Program Mentor: Dan Estes
385-432-9281 (MST)
rbuch49@wgu.edu

**Part I: Research Question**

A. Describe the purpose of this data mining report by doing the following:

1. Propose one question relevant to a real-world organizational situation that you will answer using one of the following classification methods:

   - k-nearest neighbor (KNN)
   - Naive Bayes

2. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

**A1. Proposal of Question:** Which customers are at high risk of churn? And, which customer features/variables are most significant to churn?

**A2. Defined Goal:** Stakeholders in the company will benefit by knowing, with some measure of confidence, which customers are at highest risk of churn because this will provide weight for decisions in marketing improved services to customers with these characteristics and past user experiences.

**Part II: Method Justification**

B. Explain the reasons for your chosen classification method from part A1 by doing the following:

1. Explain how the classification method you chose analyzes the selected data set. Include expected outcomes.

2. Summarize one assumption of the chosen classification method.
3. List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis.

**B1. Explanation of Classification Method:** Explain how the classification method you chose analyzes the selected data set. Include expected outcomes.

**B2. Summary of Method Assumption:** Summarize one assumption of the chosen classification method.

**B3. Packages or Libraries List:** List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis.

**Standard imports**

[ ]:

```
# Standard data science imports
import numpy as np
import pandas as pd
from pandas import Series, DataFrame

# Visualization libraries
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

# Statistics packages
import pylab
from pylab import rcParams
import statsmodels.api as sm
import statistics
from scipy import stats

# Scikit-learn
import sklearn
from sklearn import preprocessing
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import classification_report


# Import chisquare from SciPy.stats
from scipy.stats import chisquare
from scipy.stats import chi2_contingency
```

**Change color of Matplotlib font!!!**

[ ]:

```
import matplotlib as mpl

COLOR = 'white'
mpl.rcParams['text.color'] = COLOR
mpl.rcParams['axes.labelcolor'] = COLOR
mpl.rcParams['xtick.color'] = COLOR
mpl.rcParams['ytick.color'] = COLOR
```

**Ignore Warning Code**

[ ]:

```
import warnings
warnings.filterwarnings('ignore')
```

[ ]:

```
# Load data set into Pandas dataframe
df = pd.read_csv('Data/churn_clean.csv')
```

[ ]:

```
# Rename last 8 survey columns for better description of variables
df.rename(columns = {'Item1':'TimelyResponse',
                    'Item2':'Fixes',
                     'Item3':'Replacements',
                     'Item4':'Reliability',
                     'Item5':'Options',
                     'Item6':'Respectfulness',
                     'Item7':'Courteous',
                     'Item8':'Listening'},
          inplace=True)
```

**Part III: Data Preparation**

C. Perform data preparation for the chosen data set by doing the following:

1. Describe one data preprocessing goal relevant to the classification method from part A1.
2. Identify the initial dataset variables that you will use to perform the analysis for the classification question from part A1, and classify each variable as continuous or categorical.
3. Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step.
4. Provide a copy of the cleaned data set.

**C1. Data Preprocessing:** Describe one data preprocessing goal relevant to the classification method from part A1.

**C2. Dataset Variables:** Identify the initial dataset variables that you will use to perform the analysis for the classification question from part A1, and classify each variable as continuous or categorical.

**C3. Steps for Analysis:** Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step.

**C4. Cleaned Dataset:** Provide a copy of the cleaned data set.

[ ]:
```
# Create histograms of contiuous & categorical variables
df[['MonthlyCharge', 'Bandwidth_GB_Year', 'TimelyResponse', 'Courteous']].hist()
plt.savefig('churn_pyplot.jpg')
plt.tight_layout()
```

[ ]:
```
# Create Seaborn boxplots for continuous & categorical variables
sns.boxplot('MonthlyCharge', data = df)
plt.show()
```

[ ]:
```
sns.boxplot('Bandwidth_GB_Year', data = df)
plt.show()
```

[ ]:
```
sns.boxplot('TimelyResponse', data = df)
plt.show()
```

[ ]:
```
sns.boxplot('Courteous', data = df)
plt.show()
```

### Part IV: Analysis

D. Perform the data analysis and report on the results by doing the following:

1. Split the data into training and test data sets and provide the file(s).
2. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.
3. Provide the code used to perform the classification analysis from part D2.

**D1. Splitting the Data** Split the data into training and test data sets and provide the file(s).

**D2. Output & Intermediate Calculations** Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

**D3. Code Execution** Provide the code used to perform the classification analysis from part D2.

### Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:

1. Explain the accuracy and the area under the curve (AUC) of your classification model.
2. Discuss the results and implications of your classification analysis.
3. Discuss one limitation of your data analysis.
4. Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

**E1. Accuracy & AUC** Explain the accuracy and the area under the curve (AUC) of your classification model.

**E2. Results & Implications** Discuss the results and implications of your classification analysis.

**E3. Limitation** Discuss one limitation of your data analysis.

**E4. Course of Action** Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

**Part VI: Demonstration**

F. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

**F. Video** link

**G. Sources for Third-Party Code** GeeksForGeeks. (2019, July 4). Python | Visualize missing values (NaN) values using Missingno Library. GeeksForGeeks. https://www.geeksforgeeks.org/python-visualize-missing-values-nan-values-using-missingno-library/

**H. Sources** CBTNuggets. (2018, September 20). Why Data Scientists Love Python. https://www.cbtnuggets.com/blog/technology/data/why-data-scientists-love-python

Massaron, L. & Boschetti, A. (2016). Regression Analysis with Python. Packt Publishing.

[ ]:

```python
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('D209 Data Mining 1 - NVM2 - Classification Analysis.ipynb')
```

```
--2021-07-24 15:35:00--  https://raw.githubusercontent.com/brpy/colab-
pdf/master/colab_pdf.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.111.133, 185.199.109.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1864 (1.8K) [text/plain]
Saving to: colab_pdf.py

colab_pdf.py        100%[===================>]   1.82K  --.-KB/s    in 0s

2021-07-24 15:35:00 (39.6 MB/s) - colab_pdf.py saved [1864/1864]

Mounted at /content/drive/

WARNING: apt does not have a stable CLI interface. Use with caution in scripts.
```

```
WARNING: apt does not have a stable CLI interface. Use with caution in scripts.

Extracting templates from packages: 100%
```