**5-1 Assignment: Applying Predictive Models**

Kayla Sacks

Southern New Hampshire University

DAT-430 Leverage Data for Org Results

Professor Aubrey Condor

October 6, 2024

## Introduction

Predictive analytics models are used by analysts to evaluate past data, find patterns, and observe trends to predict future scenarios to help in the decision making process (Cote, 2021). There are many types of predictive models that can be used when analyzing data, each having their own use case. After explaining each model and providing an example scenario of each, I will choose two that will be for Project Two and why I chose them.

## Use Cases

Each model has its own pros and cons. It is important to choose the right model for what you need to express with the model. An example scenario for each model can help you understand if it is the right model to use to give the results they are needed from the analysis.

### Ordinary Least Squares (OLS)

The ordinary least squared method is a specific type of regression that assumes a linear relationship between the dependent and independent variables and is used to minimize the sum of squared errors to estimate the coefficients of a linear regression model (GeeksforGeeks, 2024). This method is good to use when testing relationships between variables, for example, estimating the cost of a house in specific area based on variables such as size, number of bedrooms, number of bathrooms, or age.

### Generalized Linear Models (GLM)

Generalized linear models (GLM) is a term that includes multiple models such as Linear Regression, Logistic Regression, and Poisson Regression. These models provide a way to create a linear relationship between the independent and dependent variables even if their original relationship isn’t linear (Great Learning Editorial Team, 2024). These models are adaptable and can be used when dealing with various data types. A use case for a GLM could be to determine the relationship between the number of traffic accidents and number of turns on a specific length of road such as 1, 5 or 10 mile stretches.

### Logistic Regression

Logistic regression is used to determine the probability of an event occurring and with the dependent variable being discrete, usually only having only two outcomes (Rakshith Vasudev, 2018). A use case for this type of model would be finding the probability of a customer defaulting on their credit based on their credit utilization and level of education. In this scenario, an outcome of 0 is that the customer did not default and 1 would be that the customer did default.

### Random Forests

A random forest is a collection of decision trees and can handle both classification and regression tasks. This model has improved prediction accuracy over other models and the risk of overfitting is greatly reduced since each tree is created using random subsets from the data set and measures a random subset of variables (GeeksforGeeks, 2024b). Ransom forests help to reduce the impact outliers can have as well. A random forest is a model that is good to use when trying to answer a question such as ‘does the patient have a heart disease or do they not have a heart disease?’ when using a data set that includes various health indicators such as age, resting blood pressure, maximum heart rate, and other such information.

### Decision Trees

Decision trees, like random forests, can be used when dealing with both classification and regression which has a tree structure that provides a visual outline of possible outcomes based on the input. The downside to decision trees is overfitting as it can continue to create new nodes to fit the data. A use case for this model would be when a retailer is helping a customer search for a house. They can use a decision tree to reduce the choices based on what the customer wants to find the homes they would most likely be interested in.

### Neural Networks

Neural networks are algorithms that use a process that simulates the way the human brain works to recognize the possible underlying relationships within a data set and can learn on their own once trained (Ramos, 2018). These are often used in finance but can also be used in areas such as healthcare. A well-known use case for neutral network is Google’s search algorithm. Since it is a model that continues to learn, it can also be used to help healthcare provides to make better decisions regarding treatments based information that has been gained overtime.

### Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines, or MARS, is a non-parametric regression method that automatically models nonlinearities and interactions between variables (Haleyliu, 2024). MARS models are used in many fields such as medicine, engineering, and finance. One use case for this model is predicting how many rooms will be booked at a hotel based on variables that might not have a linear connection such as the distance to other hotels, size of the hotel, or what is in the surrounding, such as stores or places to eat.

## Project Two Choices

The two models that I would apply to Project Two are logistic regression and random forest. One reason I would use these two models is because I am more familiar with them and understand that they would both work with the data set. Both models can help provide significant insights into the cause of increased attrition. As we are trying to find what conditions worsen attrition, we can use both logistic regression and random forest to find the variables that have the greatest affect. Random forest will help to prevent overfitting while improving the accuracy of predictions while logistic regression can use past data to make predictions.

## References

Cote, C. (2021, October 26). *What Is Predictive Analytics? 5 Examples | HBS Online*. Business Insights - Blog; Harvard Business School. <https://online.hbs.edu/blog/post/predictive-analytics>

GeeksforGeeks. (2024a, May 28). *Ordinary Least Squares (OLS) Regression in R*. GeeksforGeeks. <https://www.geeksforgeeks.org/ordinary-least-squares-ols-regression-in-r/>

GeeksforGeeks. (2024b, July 12). *Random forest algorithm in machine learning*. GeeksforGeeks. <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

Great Learning Editorial Team. (2024, September 2). *Generalized Linear Models | What does it mean? - Great Learning*. GreatLearning Blog: Free Resources What Matters to Shape Your Career! <https://www.mygreatlearning.com/blog/generalized-linear-models/>

Haleyliu. (2024, March 19). *Multivariate adaptive regression spline algorithm - Haleyliu - Medium*. Medium. <https://medium.com/@haleyliu388/multivariate-adaptive-regression-splines-mars-is-a-non-parametric-regression-method-that-38aead78ec3e>

Rakshith Vasudev. (2018, June 5). *How are Logistic Regression & Ordinary Least Squares Regression (Linear Regression) Related?* Towards Data Science; Towards Data Science. <https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5>

Ramos, D. (2018, October 17). *Real-Life Applications of Neural Networks*. Smartsheet. <https://www.smartsheet.com/neural-network-applications>