# Examination of Virtual Reality Experiences

## Problem Statement

Examine virtual reality experiences to improve design, comfort, and customization.

## Introduction

For the dataset relating to virtual reality experiences, I will convert the ranking system (1-5) of the immersion level into a categorical model (1-3 = Not too immersed, 4-5 = Immersed). We will then examine the response variable for immersion level through logistic regression.

In following that, we will use several predictors to demonstrate how significant of a role each one plays:

* Age
* Gender
* VR Headset type
* Duration

## Variables

**Response**: Immersion Level

**Predictor(s)**: Age, Gender, VR Headset type, Duration

## Hypothesis

**H0**: There is no relationship between the immersion level and age, gender, VR Headset type, and duration.

**H1**: There exists a relationship between the immersion level and its predictors.

## Dataset

The dataset can be found at [Kaggle.com](file:///C:\Users\atkha\Desktop\kaggle.com)

<https://www.kaggle.com/datasets/aakashjoshi123/virtual-reality-experiences>

## Motivation

I believe user experience is a useful endeavor to pursue to analyze the data through what I learned in the Machine Learning course. Products exist everywhere, and gathering user feedback and analyzing is a step of import throughout the industry.

# Project Report

## Methods

**Data Collection**: Via Kaggle using a preexisting data set for Virtual Reality Experiences.

**Data Preprocessing**: One-hot encoding. Predictors that were categorical in nature were split into separate predictors. Splits shown below:  
*Gender* – Male, Female, Other  
*VRHeadset* – HTCVive, PlayStationVR, OculusRift  
Response Variable was changed into categorical, where 1-3 was assigned to ‘NotImmersed’, and 4-5 were assigned to ‘Immersed’.

**Model Selection:** Used both Logistic Regression and Decision Trees

**Model Validation**: Tested through (10)k-fold validation and LOOCV.

**Collinearity Testing**: Tested via scatter plots.

## Results

### Model Summary

Table 1: Coefficients of each Predictor

|  |  |
| --- | --- |
| **Predictor** | **Coefficients** |
| HTCVive | -0.119 |
| PlayStationVR | -0.012 |
| OculusRift | -0.080 |
| Male | 0.024 |
| Female | -0.111 |
| Other | -0.124 |
| Age | 0.002 |
| Duration | 0.002 |

Looking at the different VR Headsets in the data, it shows that the one with the most negative correlation to the response is the HTC Vive. The PlayStation VR is very close to 0, while Oculus Rift is close to -0.1 coefficient. Gender also seems to show a negative correlation if you are not Male. The data also states that Age and Duration hardly have an affect on the model, as they are extremely close to 0.

Unfortunately, our model p-values were extremely high, making the Logistic Regression model either not be a good fit for the data set, or I incorrectly split up the data incorrectly.

Table 2: P-value of each predictor

|  |  |
| --- | --- |
| **Predictor** | **P-value** |
| HTCVive | 1 |
| PlayStationVR | 1 |
| OculusRift | 1 |
| Male | 1 |
| Female | 1 |
| Other | 1 |
| Age | 0.675 |
| Duration | 0.593 |

### Accuracy and Validation Results

For accuracy and Validation, we used the k-fold cross validation method and the Leave One Out Cross Validation (LOOCV) method. I eventually chose to use the 10-fold model, as the value of k=10 is a standard number, and the accuracy was higher in the 10-fold model rather than the 5-fold model.

Table 3: Validation Methods and Accuracy

|  |  |
| --- | --- |
| **Validation Method** | **Accuracy** |
| 5-fold | 0.601 |
| 10-fold | 0.603 |
| LOOCV | 0.603 |

While the accuracies being relatively similar looks promising, this can also demonstrate that there is linearity in the data or the data is overfit. In the case of using the Logistic Regression model, I believe that the dataset I used and categorized has cause the data to overfit. It can also be a sign that the estimate for the performance of the model is true.

### Collinearity

I used pair-plots to determine if there is a linear relationship between predictors.

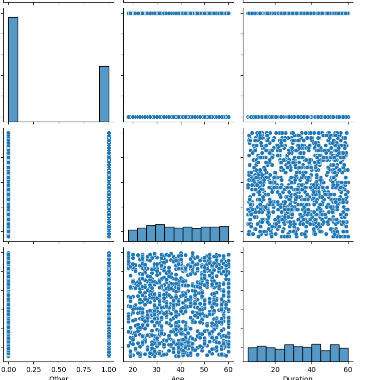


Figure 1: Snippet of pair plots for Collinearity

Due to the nature of the data being mostly categorical, we have pair plots that show has two distinct vertical lines, which can be an indication of collinearity between the predictors. There also exists plots with data only existing in one of the 4 corners of the plot. In Figure 1, we see a plot with datapoints existing all over the graph, indicating there is no collinearity between those predictors, Age and Duration. Oddly enough, Age and Duration were the ones that I did not categorize.

### PCA Results

As for our PCA results, I have solved for the values of Cn, where n is a number 1 through 8. The Eigenvalues and Eigenvectors can be seen in the code, as the list for them is huge. In Table 4, we have the coefficients listed for each Cn.

Table 5: PCs and their coefficients

|  |  |
| --- | --- |
| **PCs** | **Coefficients** |
| C1 | 0.0043 |
| C2 | -0.0259 |
| C3 | 0.0245 |
| C4 | 0.0075 |
| C5 | 0.0099 |
| C6 | -0.0086 |
| C7 | -1.008e+13 |
| C8 | 5.551e+13 |

As you can see in the table, we have C7 and C8 with coefficients in the upwards of millions, and some are zero. As I didn’t have data values remotely close to the value of millions, this can indicate that those features have a strong influence in the direction of the PC.

### Decision Tree

With all these awkward data values and misfits, I concluded that the data overfit the model. So I created a Decision Tree.

Initially, I started with a decision tree with a depth of 10. This led to an accuracy of 0.544, which is lower than the logistic regression model. I was able to get the importance of each feature.

Table 6: Feature Importance

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| HTCVive | 0.04 |
| PlayStationVR | 0.04 |
| OculusRift | 0 |
| Male | 0.02 |
| Female | 0.04 |
| Other | 0.01 |
| Age | 0.29 |
| Duration | 0.55 |

The table shows that Age and Duration are strongly important in the model of the Decision Tree. But with a relatively mediocre accuracy, I tested different depths, from 1 to 25, to check the accuracy between the test set and the training set. These values can be seen in the JupyterNotebook.

As Predicted, Training accuracy goes up, and testing accuracy goes down, which indicates our model is still overfit. Our training accuracy is much higher than our testing accuracy as the depth of the tree increases, meaning our model fits the training better than the test data.

## Future Work

I believe other models would be better to fit the dataset I have used. I also should have attempted to use KNN. With both models being overfit, It allowed me to understand how datasets overfit and no more about the limitations of both the logistic regression and Decision tree.

## References

[1] A. Joshi, “Virtual reality experiences,” Kaggle, https://www.kaggle.com/datasets/aakashjoshi123/virtual-reality-experiences/discussion/415346 (accessed 2023).