Data Science Project - Media Consumption and Internet Upgrades

Stellenbosch

UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Ruan Geldenhuys

June 2023

Full project at: https://github.com/RuanGeldenhuys/DataScience871-project

Contents

List of Figures

List of Tables

1 Introduction

The aim of this project is to investigate which factors influences an individual's willingness to upgrade their internet package, using media consumption survey data. This information can be used to employ targeting advertising techniques, focusing on people the model predicts would be willing to pay for faster internet. This ultimately decreases marketing expenditure on people who cannot be convinced to upgrade their internet speed.

I firstly, make use of exploratory data analysis and secondly, fit and tune a Random Forest model. Random Forests are a tree-based machine learning technique, that is capable of capturing complex non-linearities and generally require minimal tuning. The model in this case finds demographic factors, like age, as well as the amount of devices owned by individuals to be crucial determinants of whether an individual would pay for faster internet.

2 Data and Feature Engineering

I use survey data from the Deloitte Media Consumption survey. The data set contains survey responses from 2131 individuals regarding demographic factors (age, sex, race, ect.), the types of apps they use, time spent on different devices, preferred type of media, and other issues regarding media consumption. I restrict this data set to include only responses from individuals that have internet access. The features can be broadly broken down into 6 categories, namely demographic factors, technology owned by individuals, device usage, app usage, user subscriptions and lastly, what individuals' preferred form of entertainment are.

Binary encoding is applied to the target variable, indicating as a 1 if an individual indicated they would be willing to upgrade their internet package and a 0 otherwise. Feature variables where individuals could select multiple options, like devices owned, subscriptions purchased, etc. were handled using one hot encoding. This does lead to a large number of features, and as such increases the dimensionality of the data dramatically. After data cleaning and feature engineering, the dataset is left with 1558 observations. Of these observations, 744 individuals indicated that would be willing to upgrade their internet package. 80% of the data is used as a training set and the last 20% reserved for testing.

3 Exploratory Data Analysis

Exploratory Data Analysis is broken down into univariate- and bivariate analysis. Univariate analysis attempts to showcase the properties of the data, while bivariate analysis shows how the features relate to the target using simple graphing techniques.

3.1 Univariate Analysis

Below are breakdowns of the demographic factors of the dataset.

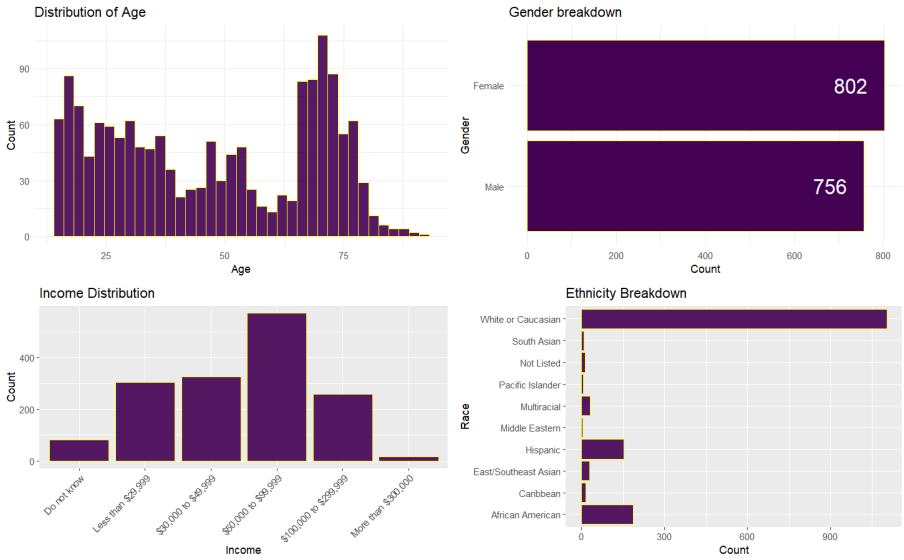

Figure 1: Demographics

A large portion of individuals who participated in the survey are elderly, as can be seen from the Age distribution that peaks around late 60's to mid 70's. The surveyed is notably lacking in middle aged respondents. Gender is split rather equally. The vast majority of respondents were white and had an income ranging from $50,000 to $99,999.

A plot of watch time, in Figure 2, of different forms of entertainment on different devices show TV's reign supreme, being the most used device to watch sport, movies and TV shows. Computer's/Laptop's are the second most used device across all entertainment, followed by smartphones and lastly tablets. This result indicates that the type of entertainment does not drastically impact the devices used.

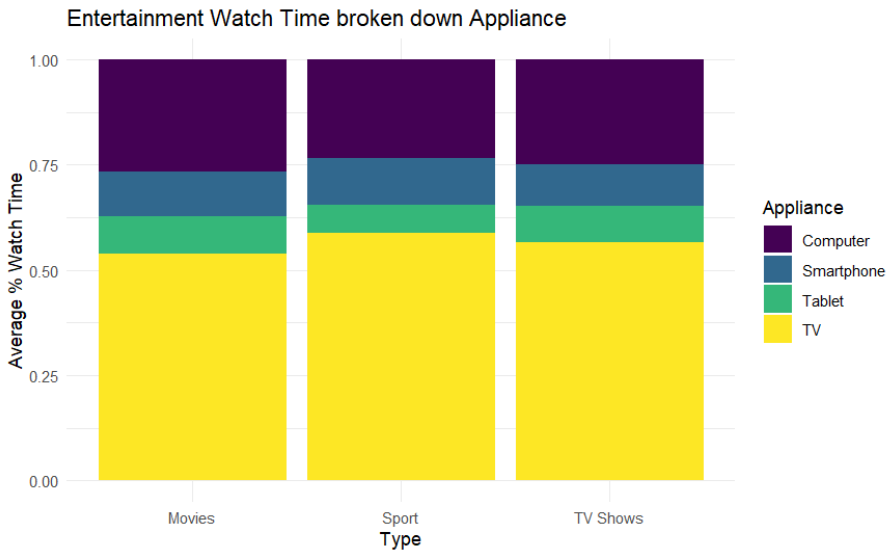

Figure 2: Watch Time

Figure 3 shows devices owned by individuals. The majority of respondents own the following:

- DVD players
- Desktop Computers
- Flat Panel Televisions
- Laptop Computers
- Smartphones
- Tablet

Plotting household subscriptions, in Figure 4, shows that most individuals are subscribed to the following services:

- Landline Telephone
- Mobile Data plan
- Mobile Voice plan
- Pay TV

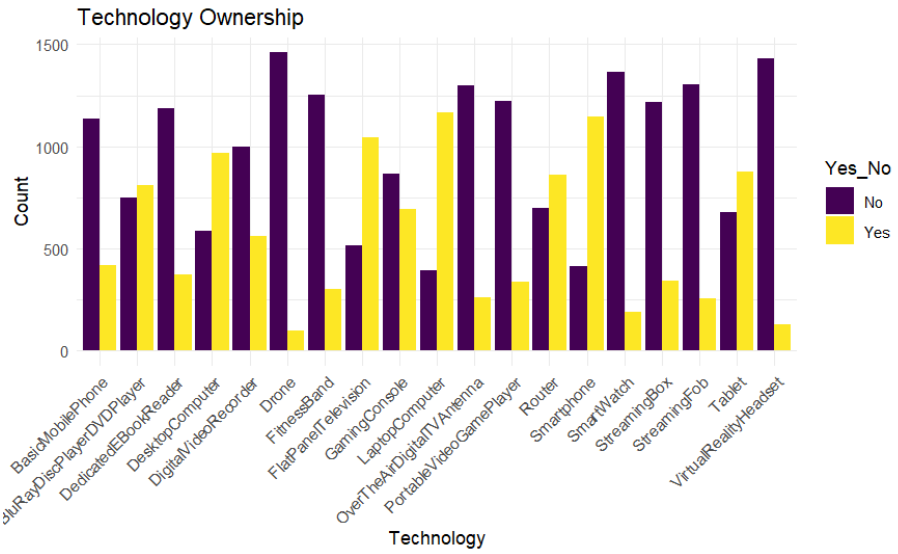

Figure 3: Technology Ownership

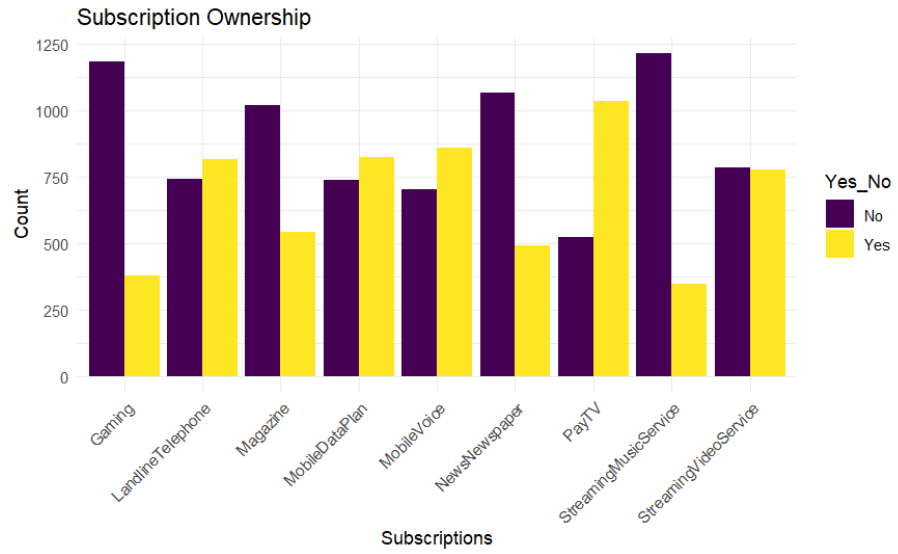

Figure 4: Subscription Ownership

3.2 Bivariate Analysis

The bivariate analysis plots several features and breaks them down based on individuals who would and wouldn't upgrade their internet package.

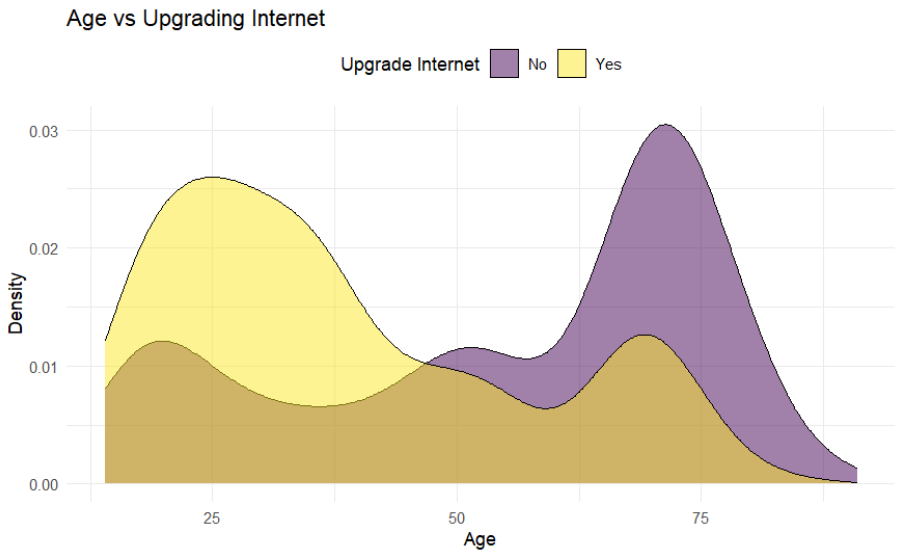

Figure 5: Age Density

What is immediately clear from Figure 5 above is the fact that younger people are far more likely to upgrade their internet speed. This peaks at around age 25. A large proportion of people who said no to upgrading their internet package are elderly.

The density plot investigating technology ownership (Figure 6) shows that the majority of people who said no, own few devices. In this survey atleast, people who own more devices are more likely to upgrade their internet speed.A similar story holds for subscriptions (Figure 7), with people who own more subscription are more likely to say yes.

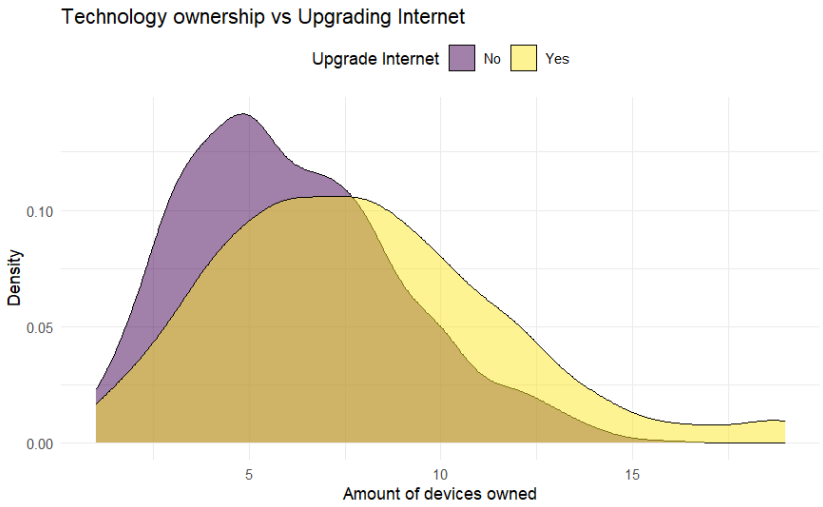

Figure 6: Technology ownership Density

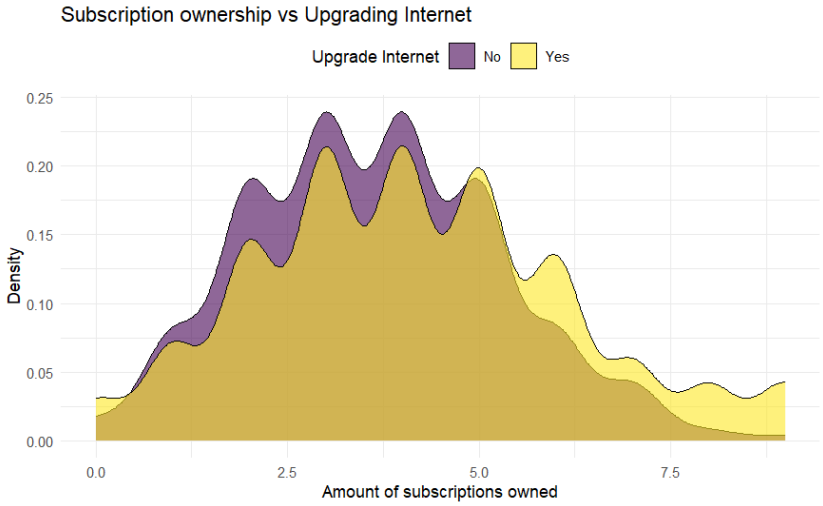

Figure 7: Subscription ownership Density

The violin charts below shows a breakdown of people who are willing to pay for upgraded internet based on different forms of entertainment and the devices used to view them.

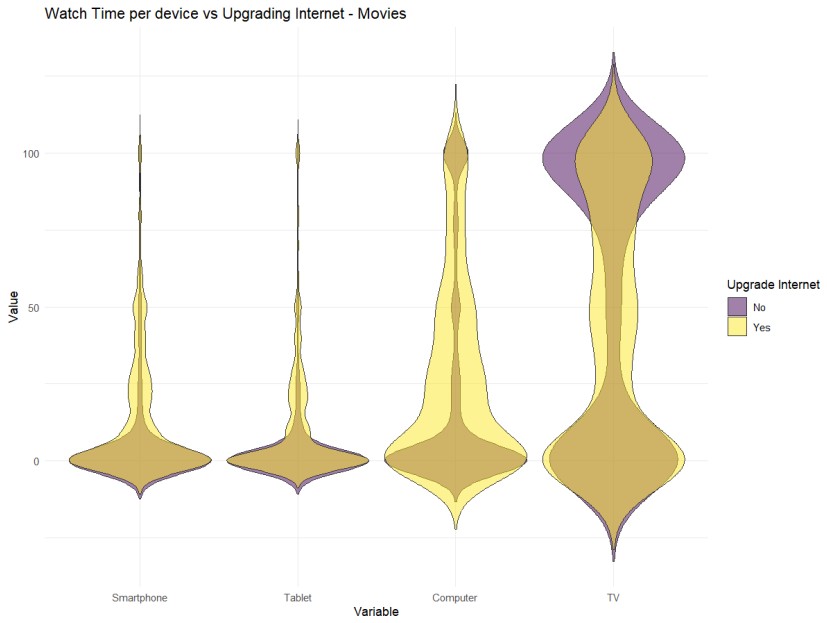

Figure 8: Movies Watch Time

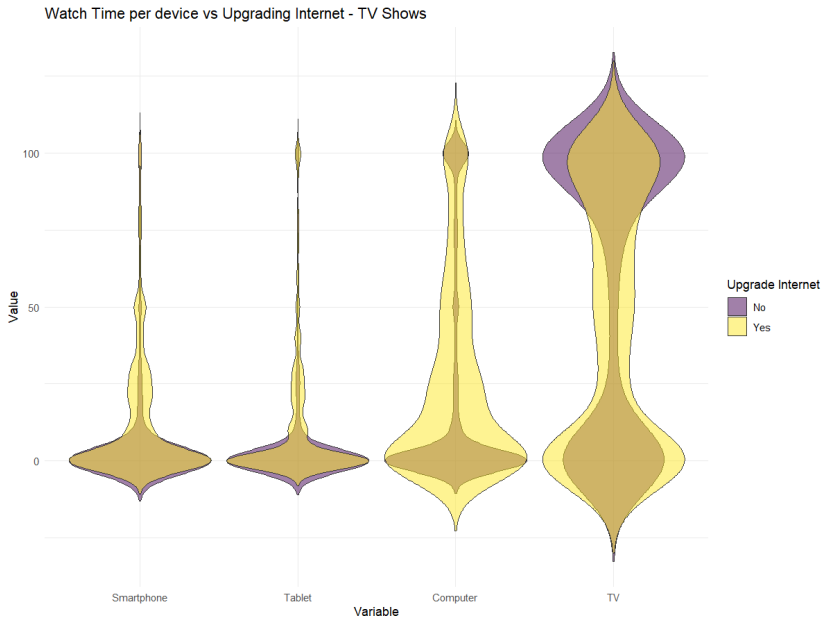

Figure 9: TV Shows Watch Time

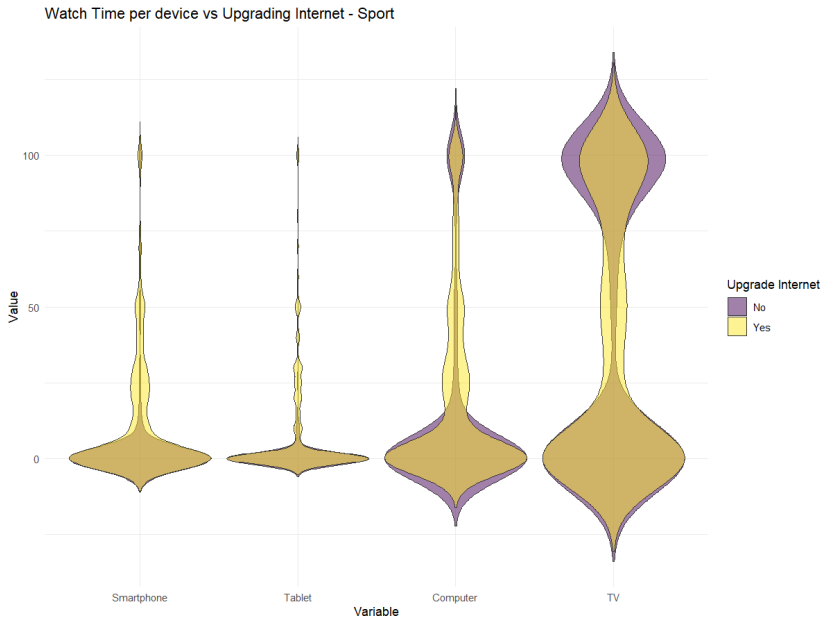

Figure 10: Sports Watch Time

Individuals who spent a significant amount of time watching Movies and TV Shows on a computer are more likely to be willing to upgrade. However, individuals who spend a large amount of time consuming any form of entertainment on TV, are much more likely to say no to an upgrade. Smartphones and tablets provide less clear results for all forms of entertainment.

4 Modeling

In order to predict which individuals are willing to upgrade their internet package I fit a Random Forest model. The target variable is the encoded as a 1 if the individual answered that they are willing to upgrade and 0 if they are not. A total of 89 features are used. The model is fitted using the ranger package.

4.1 Baseline Model

The default model in the ranger package is used as the baseline model.

The baseline model reports an Out-of-Bag (OOB) prediction error of 29% which equates to an Root Mean Squared Error (RMSE) of 0.539. This is a decent accuracy

Table 1: Baseline Model Error rates

Metric	Value
OOB Prediction Error	0.29
RMSE	0.539

Table 2: Baseline Model - Confusion Matrix (OOB)

	Predicted	
True	0	1
0	479	173
1	189	407

for a model with no tuning, which is a strength of Random Forests as stated earlier. The confusion matrix shows that the model has no preference for a certain type of mistake, with 173 false negatives and 189 false positives.

4.2 Hyperparameter Tuning

I plot baseline model Mean Squared Error (MSE) for an increasing number of trees in Figure 11 below.

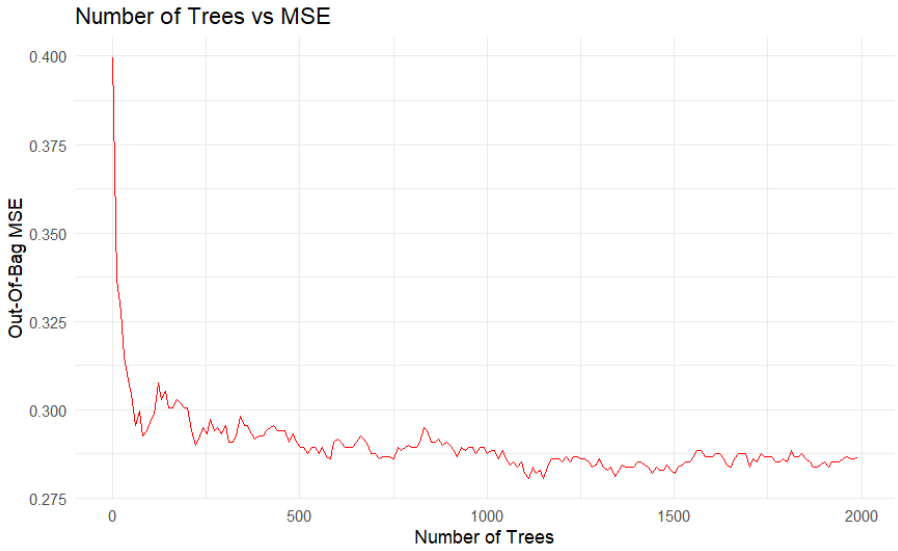

Figure 11: Baseline Model Trees vs MSE

For a small number of trees a drastic reduction in MSE is observed. This reduction seems to stabilise at around 500 trees. Although a lower MSE is possible by fitting more trees, this can cause issues. Using more trees increases the complexity of the model and thus increases the computational power required. Simply put, the reduction in errors do not justify the additional resources invested. Additionally, using a higher number of trees increases the chance that the model is overfitted, which means performance will suffer when introducing new data. I therefore continue with 500 trees in subsequent modeling.

Hyperparameter tuning was done using a standard grid search. In essence, each parameter is given a range in which the search will take place. A grid of all possible parameter combinations is created and models are iteratively fitted for every combination and the model with the lowest RMSE is selected.

Table 3: Hyperparameter Tuning Result

X	mtry	min.node.size	replace	sample.fraction	splitrule	predError	rmse
1	8	7	TRUE	1	gini	0.2732372	0.5227209

Table 3 shows the best model found. The grid search yielded only a marginal improvement, with an 0.273 OOB prediction error, which equates to an RMSE 0.523. This is a 0.016 RMSE improvement from the baseline model.

4.3 Final Model

The best model found by the grid search above is fitted.

Table 4: Final Model Error rates

Metric	Value
OOB Prediction Error	0.273
RMSE	0.523

Table 5: Final Model - Confusion Matrix (OOB)

	Predicted	
True	0	1
0	494	158
1	183	413

As stated above, this model gives marginally better predictions. The confusion matrix once again shows that the model has no preference for a certain type of mistake.

The next step is to use the model to run predictions on both the training and testing set. Note however that that the confusion matrix and error rates for the training set predictions are different from the OOB confusion matrix. This is due to the fact that OOB errors are calculated at each individual tree, based on predictions that tree makes on data not used to build it, and then aggregated across trees. Predictions across the training set, utilise all trees and all data, and merely counts the number of observations the model correctly predicted.

Table 6: Training Set Prediction Accuracy

Accuracy	0.984
Sensitivity	0.9969
Specificity	0.9698

Table 7: Training Set Prediction - Confusion Matrix

	Predicted	
True	0	1
0	650	2
1	8	578

Predictions on the training set shows an accuracy of 98.4%. Positive values are correctly predicted 99.7% of the time and negative values 96.9% of the time. However these results are not representative since the model has seen this data before. Predictions on testing data tend to be closer to the OOB accuracy and thus more representative.

Table 8: Testing Set Prediction Accuracy

Accuracy	0.7129
Sensitivity	0.711
Specificity	0.7153

The final model has a 71.29% accuracy when predicting data it has not seen before. It accurately predicts 71.11% of individuals who would be willing to upgrade their

Table 9: Testing Set Prediction - Confusion Matrix

	Predicted	
True	0	1
0	123	50
1	39	98

internet package and 71.53% of individuals who would not. The confusion matrix also shows that the model has no preference for making a certain type of mistake.

In order to interpret the model I turn the relative variable importance. Simply put, variable importance is calculated by counting how many times a variable is used for splitting a node. Variables that carry more weight in splitting the data, will appear more when selecting a variable from a subset of features at each splitting point. These results are shown in Figure 12.

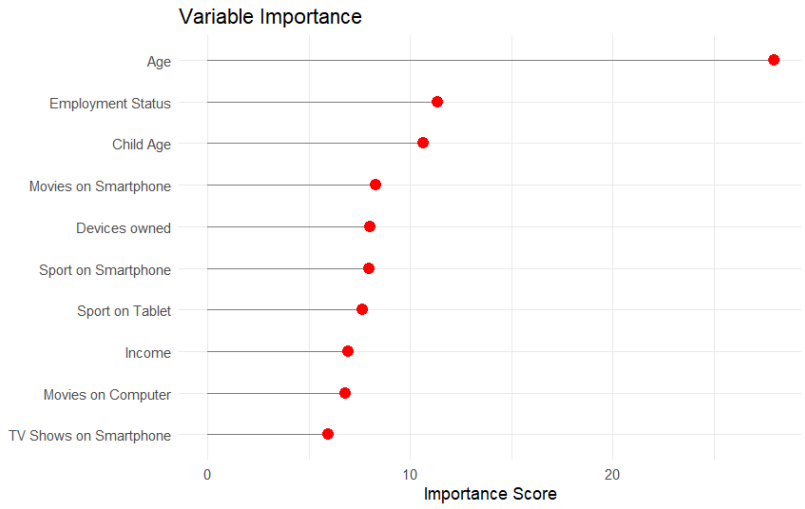

Figure 12: Variable Importance

Age is by far the most important predictor, followed by other demographic factors like child's age, employment status and income. In terms of media consumption, the amount of devices owned plays a large role. Several features involving entertainment on specific devices also show up in the top 10 features. However, in order to properly interpret these features one needs to compare them with the exploratory data analysis done earlier.

5 Discussion and Conclusion

As stated above, age seems to be the most important feature. Looking at Figure 5, its clear that young people are more likely to upgrade their internet. Thus if this model were to be used for targeted advertising, these ads should focus on people in their 20's. Technology ownership is another important feature, and as the EDA showed in figure 6, the marketing efforts should be focused on individuals with more than 8 devices in their household. As for which the devices to target, the model, together with the EDA, shows that individuals who enjoy their entertainment on smartphones, computers and tablets should be targeted.

In conclusion, the Random Forest was able to predict what individuals would be willing to upgrade their internet package with a 71.28% accuracy. The model was balanced in its errors, as it did not perform worse for a particular group. An analysis of variable importance, together with the bivariate analysis of features showed that marketing should be directed to young individuals, who own several devices and prefer to watch their entertainment on devices that are not televisions.