DELOITTE DIGITAL DEMOCRACY ANALYSIS

Who is willing to pay for faster internet?

Pranav Katariya, Masoom Patel, Max Sobkov
pkatariy, masoomp, msobkov

Contents

Foreword

Course: Data Science for Product Mangers Fall 2021

Individual Contribution

Conceptualization and Problem Formulation: Pranav, Masoom, Max
Data Cleaning and EDA: Pranav, Max
Modelling: Max, Masoom
Conclusion & Information extraction from Modelling: Pranav, Masoom, Max
Presentation: Max, Masoom, Pranav
Report: Max, Masoom, Pranav

Special thanks to Professor Steier and the TAs!

Executive Summary

PMM Consulting Inc. is working on solving a fundamental problem for Xfinity Inc. (the client) for their new product offering of lightning speed internet. The problem revolves around what kind of people will pay for their product and how they should reach out to those potential customers. The Deloitte Digital Democracy survey serves as the primary data set for analyzing these results, based on the digital consumption habits of US based customers.

Using various supervised machine learning models with a target variable of "Are customers willing to pay for lightning internet?", an accuracy of **71.04%** was obtained by performing multinomial logistic regression, one of the most important features being a "Don't know" answer to a question regarding their income bracket. Based on analysis, Xfinity should advertise to people in **income brackets of $30,000 to $100,000**, use various bundling techniques, specifically with **tablets and routers** and advertise their product on various **gaming forums**.

More specifically to these forums, when people purchase a tablet, they receive free lightning internet from Xfinity for a period of time (perhaps a month or a year). Once their trial lapses, they will be charged for the price of the internet. They should be made aware of this when signing up. The same is true for the purchase of a router. Some trial will be granted that will automatically convert into a paid subscription. Gaming forums worth advertising on are primarily gaming conventions (Electronic Entertainment Exposition and Penny Arcade Exposition), streaming platforms (Twitch.tv, YouTube Gaming, and Facebook Gaming), and potentially partnering with gaming influencers directly.

Product of Concern

Xfinity Inc.'s newly launched 'Lightning Speed' internet product offers consumers double the speed of their current internet.

Business Problem

The firm has heavy bandwidth available and needs their product to be sold as soon as possible, fearing losses by unused data. The old marketing techniques of targeting demographics and television media, flyers etc. do not seem to work for them anymore. So, there are two important questions to be answered:

1. Who are these people will pay for our lightning speed internet?

2. How should we reach out to them, so that we get maximum output and spend minimum costs on advertising vaguely?

We have tried to answer these two problems in our solution.

Client / Beneficiary

The 'firm' Xfinity Inc. benefits from the high sales of its new product offering and an indirect benefit is of the consumers who are in real need of this product and are accurately found out by Xfinity Inc. to be the probable customers.

The Dataset

The primary dataset is the Deloitte Digital Democracy Survey. Specifically, results from the 2010 and 2011 years of running the survey.

There are approximately 25 questions on Digital Consumption Lifestyle and Demographics.

Demographics

1. Age
2. Gender
3. US State
4. Region
5. Ethnicity etc.

Digital Consumption

For the model, we are considering only the following questions from the survey:

Which of the following media or home entertainment equipment does your household own?
Please select all that apply.

Row:

[r1] Flat panel television

[r2] Digital video recorder (DVR)

[r3] Streaming media box or over-the-top box

[r4] Portable streaming thumb drive/fob

[rNew1] Over-the-air digital TV antenna (for free access to network broadcast without pay TV subscription)

[r5] Blu-ray disc player/DVD player

[r6] Gaming console

[r7] Portable video game player

[r8] Computer network/router in your home for wireless computer/laptop usage

[r9] Desktop computer

[r10] Laptop computer

[r12] Tablet

[r14] Dedicated e-book reader

[r15] Smartphone

[r17] Basic mobile phone (not a smartphone)

Of those products you indicated you do not currently own, which of the following do you plan to purchase in the next 12 months? [only served up those products that consumers indicated they do not own]
Please select all that apply.

Row:

[r1] Flat panel television

[r2] Digital video recorder (DVR)

[r3] Streaming media box or over-the-top box

[r4] Portable streaming thumb drive/fob

[rNew1] Over-the-air digital TV antenna (for free access to network broadcast without pay TV subscription)

[r5] Blu-ray disc player/DVD player

[r6] Gaming console

[r7] Portable video game player

[r8] Computer network/router in your home for wireless computer/laptop usage

[r9] Desktop computer

[r10] Laptop computer

[r12] Tablet

[r14] Dedicated e-book reader

[r15] Smartphone

[r17] Basic mobile phone (not a smartphone)

[r18] Smart watch

[r19] Fitness band

[rNew2] Virtual reality headset

[rNew3] Drone

[r22] None of the above

Of the time you spend watching movies, what percentage of time do you watch on the following devices?

Row:

[r1] Smartphone

[r2] Tablet

[r3] Laptop/Desktop

[r4] Television

[r5] I do not watch movies [EXCLUSIVE ANSWER]

Of the time you spend watching TV shows, what percentage of time do you watch on the following devices?

Row:

[r1] Smartphone

[r2] Tablet

[r3] Laptop/Desktop

[r4] Television

[r5] I do not watch TV shows [EXCLUSIVE ANSWER]

Which of the following subscriptions does your household purchase?

Please select all that apply.

Row:

[r1] Pay TV (traditional cable and/or satellite bundle)

[r2] Home internet

[r3] Landline telephone

[r4] Mobile voice (smartphone or basic mobile phone calling plan)

[r5] Mobile data plan

[r6] Streaming video service

[r7] Streaming music service

[r8] Gaming

[r9] News/Newspaper (print or digital)

[r10] Magazine (print or digital)

[r11] None of the above

Of the services you indicated your household purchases, which [totalcount] do you value the most?

Row:

[r1] Pay TV (traditional cable and/or satellite bundle)

[r2] Home internet

[r3] Landline telephone

[r4] Mobile voice

[r5] Mobile data plan

[r6] Streaming video service

[r7] Streaming music service

[r8] Gaming

[r9] News/Newspaper (print or digital)

[r10] Magazine (print or digital)

You said that you subscribe to home Internet access, how much more would you be willing to pay to receive double your download speed?

Please select one.

Row:

[r1] I am willing to pay $5 per month on top of what I already pay

[r2] I am willing to pay $10 per month on top of what I already pay

[r3] I am willing to pay $20 per month on top of what I already pay

[r4] I am willing to pay $30 or more per month on top of what I already pay

[r5] I am not willing to pay more for faster download speeds as my current speed is sufficient for my needs

[r6] I prefer faster speed, but I am unwilling to pay more than I already do

Do you ever "binge-watch" television shows, meaning watching three or more episodes of a TV series in one sitting?

Row:

[r1] Yes

[r2] No

Data Cleaning

1. Feature selection

```
In [8]: # Filter data sets on questions we believe are helpful predictors
        regex_string = r'(.*Q29.*)|(.*QNEW24.*)|(.*Q37r10.*)|(.*Q73r13.*)|(.*QNEW3.*)|(.*Q6.*)|(.*(Q8|Q10).*(-Tablet|-Streami

In [9]: pd.set_option("display.max_columns", None)

In [10]: dfnew2 = df2.filter(regex=regex_string)
         dfnew2

Out[10]:
```

Q8 - Which of the following media or home entertainment equipment does your household own?-Desktop	Q8 - Which of the following media or home entertainment equipment does your household own?-Laptop computer	Q8 - Which of the following media or home entertainment equipment does your household own?-Tablet	Q8 - Which of the following media or home entertainment equipment does your household own?-Smartphone	Q8 - Which of the following media or home entertainment equipment does your household own?-None of the above	Q10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase in the	Q10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase in the	Q10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase	Q10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase in the	Q10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase in the next 12 months? -Computer network/router in your home for	Q10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase	Q10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase	(p ir c v fc

2. Null value transformation

23% of the data set was dropped due to null values. 3345 rows remain which is a very significant amount of data still, so this is a valid treatment of null values.

```
# Shape before and after dropping rows with missing values
print(df_merged.shape)
df_merged.dropna(inplace=True)
print(df_merged.shape)
```

```
(4340, 49)
(3345, 49)
```

3. One-hot encoding

Columns which contained multiple answers were encoded as a 1 for belonging to that group and a 0 for not belonging to that group. A user can only fall into a single income bracket, so this methodology is perfect for transforming categorical data into numerical data since a value can only be 1 or 0

Example of transformed income data:

nt	Q6 - Into which of the following categories does your total annual household income fall before taxes? Again, we promise to keep this, and all your answers, completely confidential.	0	1	2	3	4	5	6
0	50, 000to 99,999	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0	30, 000to 49,999	0.0	1.0	0.0	0.0	0.0	0.0	0.0
0	100, 000to 299,999	1.0	0.0	0.0	0.0	0.0	0.0	0.0

4. Binary Encoding

Values which were stored as yes/no were converted into 1/0 for modeling

	Q8 - Which of the following media or home entertainment equipment does your household own?-Streaming media box or over-the-top box	Q8 - Which of the following media or home entertainment equipment does your household own?-Portable streaming thumb drive/fob	Q8 - Which of the following media or home entertainment equipment does your household own?-Gaming console	Q8 - Which of the following media or home entertainment equipment does your household own?-Portable video game player	Q8 - Which of the following media or home entertainment equipment does your household own?-Computer network/router in your home for wireless computer/laptop usage	Q8 - Which of the following media or home entertainment equipment does your household own?-Desktop computer	Q8 - Which of the following media or home entertainment equipment does your household own?-Laptop computer	Q8 - Which of the following media or home entertainment equipment does your household own?-Tablet	Q8 - Which of the following media or home entertainment equipment does your household own?-Smartphone	Q8 - Which of the following media or home entertainment equipment does your household own?-None of the above
0	0	0	1	0	0	1	1	1	1	0
1	0	0	1	0	1	0	1	1	1	0
2	0	0	0	1	1	1	1	1	1	0

Modelling Approach

Modelling was performed using various supervised machine learning models, with a target variable of "Is the user interested in paying for lightning internet?" The data was split into 90% training data and 10% testing data. A fixed random seed was used to ensure no p-hacking took place.

A decision tree was used first since it is very explainable usually. The hope was that it would be an effective solution

Decision Tree

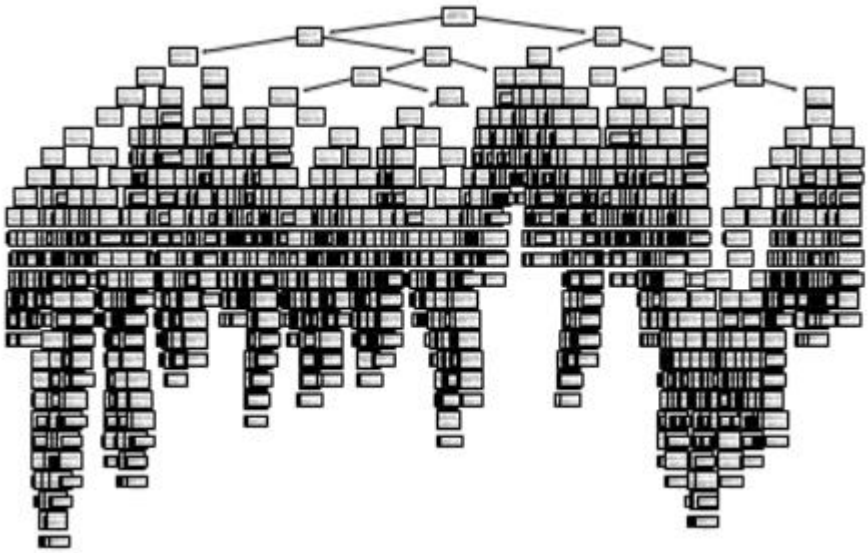
Decision Tree

In [32]:
```
# Building Decision Tree
dt = DecisionTreeClassifier(criterion = 'entropy', random_state = 42)
dt.fit(X_train, Y_train)
dt_pred_train = dt.predict(X_train)
```

In [33]:
```
# Building Decision Tree (Max depth limited)
dt = DecisionTreeClassifier(criterion = 'entropy', random_state = 42, max_depth=5)
dt.fit(X_train, Y_train)
dt_pred_train = dt.predict(X_train)

figure(figsize=(50,50),dpi=100)
tree.plot_tree(dt)
plt.show()
```

Accuracy: 52.67%



Unused, but attempted models

Random Forest: - 49.25% accuracy
Support Vector Machines - 68.96% accuracy
K-Nearest Neighbors - 66.87% accuracy
Gradient Boosting - 69.85% accuracy
Naive Bayes - 69.55% accuracy

Logistic Regression

```
# Rerun Logistic Regression on most important features
X=df_merged[coefs["Questions"]].values
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.1, random_state = 42)
clf2 = LogisticRegression(max_iter=1000, random_state=42).fit(X_train, Y_train)
print("Test accuracy: " + str(clf2.score(X_test, Y_test)))
```

```
Test accuracy: 0.7104477611940299
```

Accuracy: 71.04%

Logistic Regression Modelling Reflection:

Logistic regression is a great model, because it is trivial to find how important each feature is to the overall outcome, which makes for better recommendations than a black box model. It also specializes in classification of data, particularly for the purposes of binary classification in its base state.

Confusion matrix:

```
y_pred = clf.predict(X_test)
y_actual = Y_test

tn, fp, fn, tp = confusion_matrix(y_actual, y_pred).ravel()
```

```
print("Test true negatives: " + str(tn))
print("Test true positives: " + str(tp))
print("Test false negatives: " + str(fn))
print("Test false positives: " + str(fp))
```

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	49	39
Predicted Negative (0)	62	185

A preferable model would have more false positives (Predicted 1 Actually 0) than false negatives since it is better to advertise to everyone who wants faster internet and potentially advertise to some who don't, than miss out on a single potential customer.

Features that could potentially improve the model:

Knowing a user's internet usage over a fixed time period (perhaps a month)

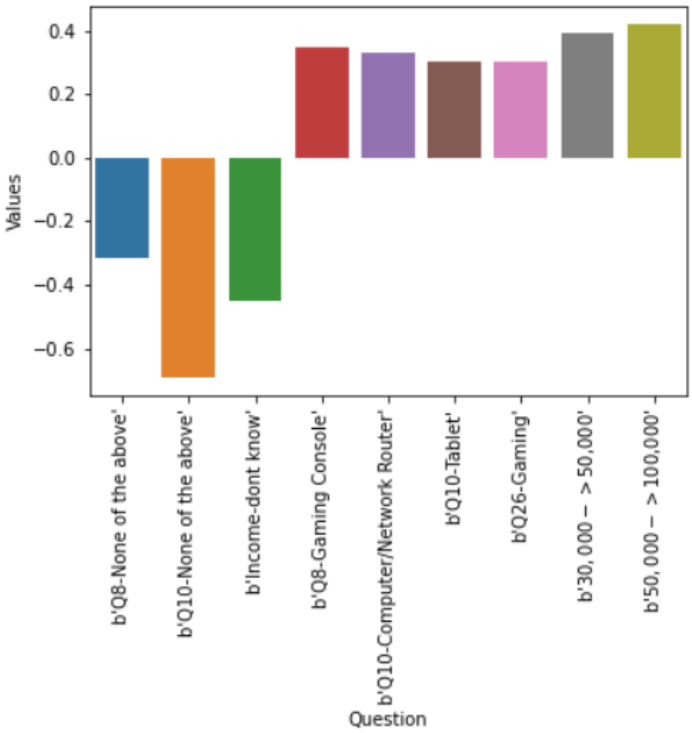
Rationale: Users with higher internet usage would benefit more from lightning internet

Advertise directly to companies with large work from home programs

Rationale: Companies are a smaller group to sell to and they may subsidize employee's internet, so partnering directly with them could indirectly gather many more "customers"

Top 9 most important predictors:

1. Income
   Answer: None of the Above
2. Devices you are planning to buy in the next 12 months
   Answer: None
3. Which of the following media/entertainment devices does your household own?
   Answer: None
4. Which entertainment device does your household own?
   Answer: Gaming Console
5. Which of the brackets does your annual household income fall in?
   Answer: $30,000 to 50,000 and $50,000 to $100,000
6. Devices you are planning to buy in the next 12 months
   Answer: Tablet or Router



Recommendations

Based on the model predictions and after extracting the most important features, the following recommendations represent the findings of the consulting team:

1. Do not advertise to people with annual incomes above $100,000 as they own the highest possible speed packages already.
2. Advertise to the people falling into the annual income bracket of $30,000 to $100,000 as these are the customers most likely to buy upgraded internet plans.
3. Offer bundled packages of internet with various Tablet sales. Once their trial subscription runs out, they will be attached to the faster internet, and will be converted into paying customers.
4. Offer bundling to people who are looking to buy new routers, since a faster router can be bottlenecked by poor internet service.
5. Advertise on various gaming forums, or platforms as people who are using the gaming devices require high speed computing and will be willing pay more for higher speed.

Reference

[1] Deloitte Insights. “Digital Media Trends, 15th Edition.” Accessed October 13, 2021. https://www2.deloitte.com/us/en/insights/industry/technology/digital-media-trends-consumption-habits-survey.html.