Making Products Count: Data Science for Product Management

95851, Section B3

Spring 2022

Uncovering and adopting untapped streaming consumers

Project Report, Option 2, Survey Analysis

Professor David Steier

March 4, 2022

Nicholas Thomas, nhthomas

Kelly McManus, kellymcm

Davidson Siga , dsiga

Krishnaraj Pawooskar, kpawoosk

Michael Yang, myang3

Carnegie Mellon University

Heinz College

Table of Contents

Executive Summary

The streaming industry has been growing year over year, attracting new entrants and creating intense competition. Streaming companies are competing for consumer attention and subscriptions. This has led to an increase in the cost of acquisition. To mitigate this problem, the earmark team has developed a predictive model using surveys and public data to identify if an individual is a current streaming consumer or a high potential lead. The model allows streaming companies an opportunity to increase their return on ad spend and overall market share. This report outlines the process of building this predictive model, including exploratory data analysis, data preparation, model selection, evaluation, and recommendations.

Introduction

Problem

In today's increasingly digital world, consumers are moving online for sources of entertainment. In particular, the video streaming market has seen an increase in revenue and new entrants since 2010. The US video streaming market alone is predicted to increase at an annual rate of 23.2% for the next five years and reach annual revenue of $119 billion (Cook, 2020). New enterprises are looking to join this attractive market, but are in need of a way to acquire new customers and upsell customers to premium subscriptions. Streaming companies are faced with the challenge of many consumers subscribing to competitors and being unwilling to add an additional streaming service to their subscription expenses (Cook, 2020). For these reasons, streaming companies require tools to identify high potential leads so they can better advertise their products and premium services to increase customer conversions and ROI on advertising, while also reducing the cost of acquisition.

As experienced product managers and data scientists, our team came together to found *earmark*, a business intelligence firm that helps streaming companies uncover and adopt untapped streaming consumers and achieve growth.

Benefits

New technology and innovation are at the center of the streaming industry. Companies that adopt innovative approaches to business are the ones that will survive potential industry consolidation. Earmark's prediction algorithms will be a competitive advantage to streaming companies and will allow them to predict consumers who are already streaming subscribers and focus their advertising effort to target those that are yet to adopt a streaming service.

Definitions

In this study, **streaming services** are defined as the presence of a streaming music service (Spotify, Apple Music), streaming video service (Netflix, Amazon Video), or gaming subscription (Steam). A respondent who purchases one of these services is a **streamer**.

Goal

For the purpose of this study, earmark will walk through exploratory data analysis, data preparation, feature engineering, model building, and evaluation. The first iteration of our prediction model answers the question: Is this user already a streaming subscriber?

Exploratory Data Analysis

Data Source Overview

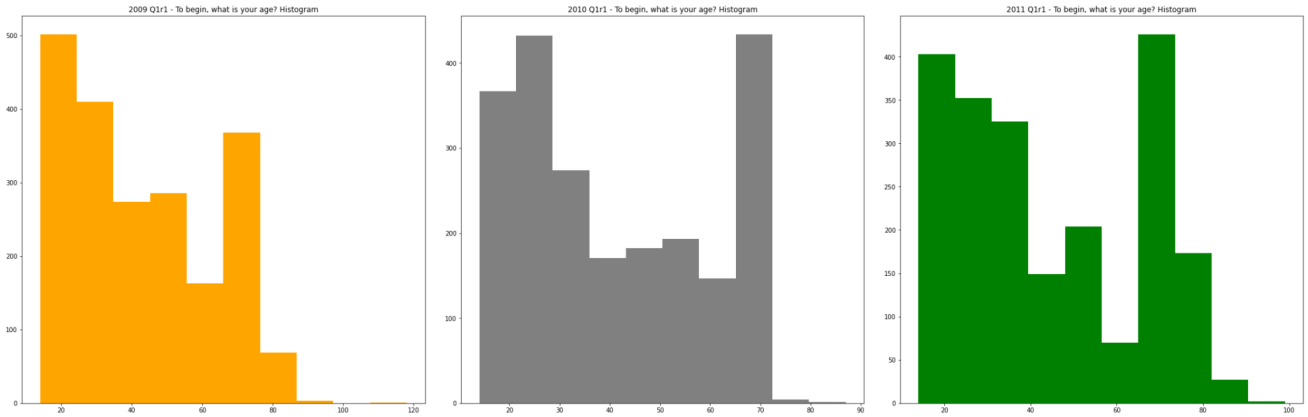
Our models utilize survey data around media consumption habits gathered by Deloitte in preparation for their annual Digital Democracy report. This data contains demographic information such as age, ethnicity, location, and income as well as information about the frequency, distribution channels, and devices respondents leverage to consume media. There are three datasets for the years 2009, 2010, and 2011. Earmark combined the various years of respondent data to build our predictive models.

Demographic Information

The respondents had a nearly 50-50 split between gender (male/female) for each of the years of respondent data.

Looking at age, the majority of the respondents skewed younger. Notably, there were many other respondents between the ages of 70-80. When combining the three datasets, the distributions did not change significantly as there was relatively equal distribution as seen below:

Figure 1 - Frequency distribution of Age for datasets 2009, 2010, and 2011



Another factor to consider is ethnicity. Table 1 compares the ethnicity split for the United States, provided by census.gov, and the Deloitte dataset.

Table 1 - Ethnicity, United States vs. Survey Data

Ethnicity	census.gov	Deloitte Data
White	76.3%	64-70%
Hispanic	18.5%	10-11%
Black / African American	13.4%	10-13%
Asian	5.9%	2-7%

While not an exact match, the Deloitte dataset closely resembles the composition of the United States. An important consideration for any company that seeks to model populations is that if this model is applied to an audience that doesn't represent this composition, the effectiveness of an ad campaign could suffer.

The respondent's incomes primarily fell into 3 categories:

- Less than $29,999 (~25%)
- $30,000 to $49,999 (~20%)
- $50,000 to $99,999 (~30%)
- $100,000 to $299,999 (~15%)

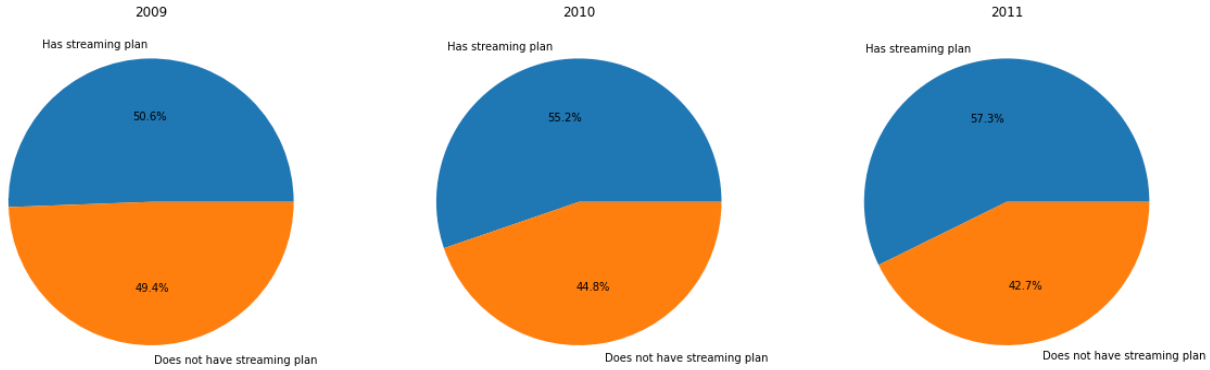
There were very few wealthy individuals making more than $300,000 and there was another small population that answered "Do not know". For the respondents who did not know their income we binned this population in with the $50,000 to $99,999 group as the median income for the United States is approximately $68,000 according to census.gov.

Media Consumption Trends

Analyzing the media consumption questions of the survey, streaming is identified as the most frequent entertainment used by consumers, but it is not valued as highly as pay-TV, internet, or phone as indicated in Appendix A and B. This shows that streaming services should work to make their products an essential part of consumers' lives by including content like self-serve news, media biographies, educational programs.

Considering the overall trend of purchasing streaming services, year over year, an increasing percentage of consumers are paying for streaming services, as seen in Figure 2. This growth in the streaming population makes the market attractive for new entrants. However, streaming companies also need to be wary that the growing streaming population creates competition for gaining consumer attention and subscriptions.

Figure 2 - Has vs. does not have a streaming plan, 2009-2011



Response Variables

Now having a grasp of our population's background and entertainment consumption trends, we need to understand how our response variables are distributed. Our primary model output is a prediction of whether or not a person is a streamer. We use owning streaming equipment as the indicator that they are or are not a streamer. If the person does not have streaming equipment, our business can inform decisions about targeting ads to these users. Overall, we see that a slim majority of consumers have a streaming plan. This presents a target-rich environment and opportunity to convert consumers to streaming platforms.

Data Preparation

It is often said in data science, garbage in produces garbage out. While methods such as autoML and deep feature synthesis can automate much of the modeling workflow, parsimony and interpretability are also important considerations. Cleaning the data and engineering features allows not only us but also our clients to gain additional predictive value from the data we have on tap. At earmark, we carefully groom and engineer features to boost our supervised models' predictive power.

Data Leakage

Just like a leaky water pipe can ruin the inside of your house, data leakage can spoil any data science application. Data leakage occurs when a model has access to information that it should not. In earmark's case, the Deloitte survey data contains information that would immediately indicate the respondent is a streaming customer. For instance, the survey asks respondents if they use a streaming media box, over-the-top box portable streaming thumb drive/fob, or watch digital video entertainment via an online streaming service. These features leak information into a model that should not be there. Earmark excluded similar columns to ensure the model generalizes well and provides value for our clients.

Feature Engineering

As seen in the exploratory data analysis, the dataset is rich, combining three years of survey data and 195 features to build our models. Before modeling, ordinal, categorical, and numerical features need to be prepared.

Out of this plethora of data, earmark extracted several features that proved invaluable in predicting if a customer would buy a streaming service. Namely, we targeted feature engineering on six survey questions shown in Table 2.

Table 2 - Feature Engineering Questions

Question Number	Description	Ending Transformation
Combined Question 1&2	Binned categorical information on the age of children in a home	Distilled to 3 binary columns: has children, has children under 18, has children over 18
Question 3	Categorical data on employment status (student, retired, unemployed, employed)	Extracted binary flag for student
Question 6	Categorical data on respondents' income	Bucketed data into lower income, middle income, and upper-income ranges
Question 8	Binary data on the presence of technological devices in the home (streaming box, drone, flatscreen tv, etc.)	We created a binary flag for the presence of a “high-tech person” who owns either a smartwatch, fitness band, virtual reality headset, or drone. This flag indicates that the person is most likely tech-enabled and would likely have a streaming service using business logic We created a binary flag for the presence of a “low-tech person” who owns either an over-the-air antenna or a basic mobile phone (not a smartphone). These are features of a person that usually isn’t technology-savvy and would most likely not have a streaming service
Question 26	Question about services that a household pays for	This question served as our target variable for one of our model outputs. We defined streaming as the presence of streaming video, streaming music, or gaming subscription as described in the exploratory data analysis section

Furthermore, there were specific types of transformations that we conducted on the features: binary response, numerical range, and multi-categorical.

Binary & Multi-Categorical Response

For binary or multi-categorical data and columns with Yes/No values, we use one-hot encoding to transform their data type to numerical. Therefore, each category will have its own column with only the value of 0 and 1. If the column value of a record is 1, the record belongs to such a category. We transformed the binary responses for questions 1, 2, 3, 8, 26 & 37 to have numerical 0 (No) or 1 (Yes) values after bucketing. An example of these transformations are seen in Figures 3 and 4.

Figure 3 - Questions before binary cleaning

QNEW1 - Children	QNEW2 - Children, 0-4	QNEW2 - Children, 5-9	QNEW2 - Children, 10-13	QNEW2 - Children, 14-18	QNEW2 - Children, 19-25	QNEW2 - Children, 26+
No	NaN	NaN	NaN	NaN	NaN	NaN
No	No	Yes	No	No	No	No
No	No	No	Yes	Yes	No	No
Yes	NaN	NaN	NaN	NaN	NaN	NaN
Yes	No	Yes	No	No	No	No

Figure 4 - Questions after binary cleaning

QNEW1 - Children	QNEW2 - Children, 0-18	QNEW2 - Children, 19+
0	0	0
0	1	0
0	1	0
1	0	0
1	1	0

Numerical Range Response

For responses that indicated a bucketed range of choices, we converted these values to larger buckets with binary values. A 1 indicates the respondent belongs in the column bucket.

Figure 5 - Question before numerical range cleaning

Q6 - Income
50,000to99,999
50,000to99,999
50,000to99,999
100,000to299,999
Less than $29,999

Figure 6 - Question after numerical range cleaning

Q6 - Income, LowerClass	Q6 - Income, MiddleClass	Q6 - Income, UpperClass
0	1	0
0	1	0
0	1	0
0	0	1
1	0	0

Ordinal Response

Another thing to be noted is that some features with categorical data, like questions 29 & 39, have ordinal answers. Therefore we assigned numerical values to these fields. For question 29, we extracted the number from the choices and made them the new values of that column as seen in Figures 7 and 8. For question 39, we assigned a range of numbers from 1 to 4 as the degree of the action.

Figure 7 - Questions before data extraction

Q29 - Willing to pay, Speed
I am willing to pay $10 per month on top of what I already pay
NaN
I am willing to pay $20 per month on top of what I already pay
I am not willing to pay more for faster download speeds as my current speed is sufficient for my needs
I prefer faster speed but I am unwilling to pay more than I already do

Figure 8 - Questions after data extraction

Q29 - Willing to pay, Speed
10.0
NaN
20.0
0.0
0.0

Data Cleaning

Certain data types of input are not accepted for some models. In order to examine the feature importance and test different models in our hypothesis space, it is crucial to have numerical,

clean values. We performed outlier detection and imputed missing values to fix these concerns and eliminate noise.

Outliers

After creating impactful features through feature engineering, we are able to conduct outlier detection and handle any discovered outliers. First, the summary statistics are computed to observe any outliers that need to be corrected. Outliers were identified in columns seen in Table 3.

Table 3 - Outliers in Dataset

Feature	**Outlier Handling**
Age	The maximum age in the data set is 118, an unrealistic value. To fix this, we found the age at the 99th percentile of the dataset, which was 79 years old. The top 1 percent of data points were dropped and data corresponding to age < 80 remained.
Q15, Q16, Q17	These questions indicated the percentage of time respondents used devices to view various content. These 3 questions had a total of 12 columns. The minimum value of these columns was -1. The following steps were taken to handle these outliers: ● Rows with more than eight occurrences of -1 values were removed. This indicated user error or a disinterested user filling the survey. ● Rows with fewer than eight occurrences of -1 had the value overwritten by 0.

Missing Values

The dataset also suffered from missing values. For question “Q29 - Willing to pay, spend”, we imputed missing values with the column mean. For the full dataset, we imputed “NA” values with 0. This remedies all columns containing missing values.

Model Selection

Earmark believes in starting simple and gradually increasing complexity in processes for a safe and sustainable solution for our customers. In today’s increasingly digital world, Earmark’s primary goal is to successfully identify the digital streaming subscriber. Identifying this key characteristic will help earmark assist streaming industry players like Netflix, Amazon Prime, Hulu, etc. to improve product customization and ad targeting for a consumer’s best experience.

We value best product management practices assisted with state-of-the-art data science solutions. Firstly, our product managers identify the problem and work with our data scientists to determine statistical and business best case practices to identify the key data features which can help us solve the problem at hand. Rigorous, yet information-preserving data transformation techniques are performed to get the best information for modeling. Below are the details of Model Selection and what goes into it.

Problem Statement

We would like to classify a customer as being a streamer or not. The outcome variable is a 1 if they are a streamer and a 0 if they are not a streamer.

The Approach

We use the industry-standard data science pipeline that follows the steps of obtaining data, scrubbing data, exploring data, modeling, and interpreting the results to solve this problem and predict the outcome variable (Lau, 2019).

Obtain

Publicly available three years of Deloitte's reliable data sources are imported to our AI Platforms with an aim to bring structure, meaning, and insight into the data. The Python libraries Pandas, NumPy, Matplotlib, and Seaborn are used to obtain and explore the data effectively.

Pandas is a fast, powerful, flexible, and easy-to-use open-source data analysis and manipulation tool, built on top of the Python programming language.

NumPy is the fundamental package for scientific computing with Python.

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Figure 9 shows a selection of the data.

Figure 9 - Data selection

	Age	Age - Bucket	BingeWatcher	Employment - Employed or not	Employment - Student or not	Ethnicity	Gender	Potential Premium Customer	Q10 - Media equipment, Plan to purchase, 3D printer	Q10 - Media equipment, Plan to purchase, Basic mobile phone	eq p
0	28.0	24-29	0	0	0	South Asian (India, Pakistan, Sri Lanka)	Female	1	No	No	
1	33.0	30-46	0	0	0	White or Caucasian (Non-Hispanic)	Female	0	No	No	
2	24.0	24-29	1	1	0	White or Caucasian (Non-Hispanic)	Male	1	No	No	

Scrub and Explore

Data cleaning and preparation is a complex process that if done incorrectly can lead to loss of important information from critical data. The Data Preparation section described in detail the steps we took to prepare the data for modeling. Similarly, data exploration was conducted prior to modeling and is discussed in Exploratory Data Analysis.

Model

Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience (Mitchel, 1997).

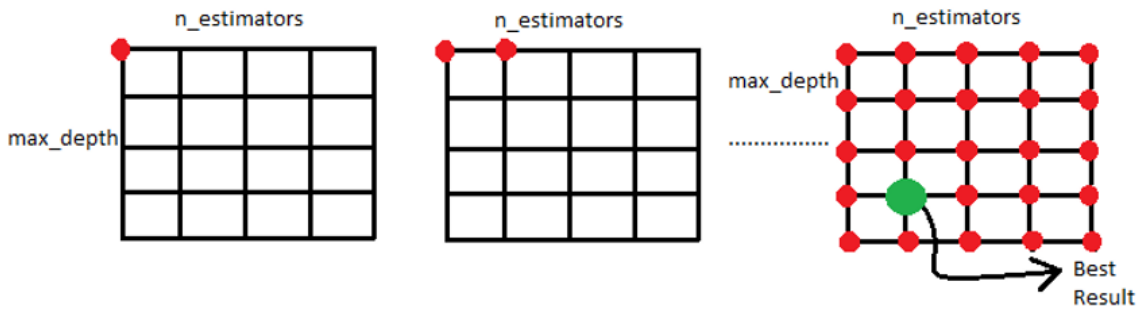
We choose to start simple with logistic regression and then move into a higher complexity-based random forest model to solve the problem statement of if an individual is a streamer or not. This is to ensure an iterative understanding of the ability of classifiers to predict information. (Akella, 2022)

Logistic regression is a statistical model used to determine if an independent variable has an effect on a binary dependent variable(DeepAI, n.d.). This means that there are only two potential outcomes given an input.

Random forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification (Wood, n.d.). It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

Using grid search, we scan the data to configure optimal parameters for a given model (Lutins, 2017). Depending on the type of model utilized, certain parameters are necessary. Grid search does not only apply to one model type. Grid search can be applied across machine learning to get the best of any model for the best predictions. An example of the grid search process is seen in Figure 10.

Figure 10 - Grid Search (Munagala, 2021)



When random forest and logistic regression classifiers were run with the grid search, we found the best results in the random forest model.

Evaluation & Interpretation

It is essential to produce high performance reliable, and interpretable results for our product to deploy for stakeholders.

A classification problem is often judged through a variety of metrics depending on the context of the business problem. The common performance measures for classifier models are accuracy, F1 score, precision, recall, ROC curves, and AUC.

We believe the cost of predicting incorrectly is much higher than that of the cost of predicting correctly because we don't want stakeholders to invest in a product that might lead to unnecessary expenditure for them. We will consider this when evaluating the results.

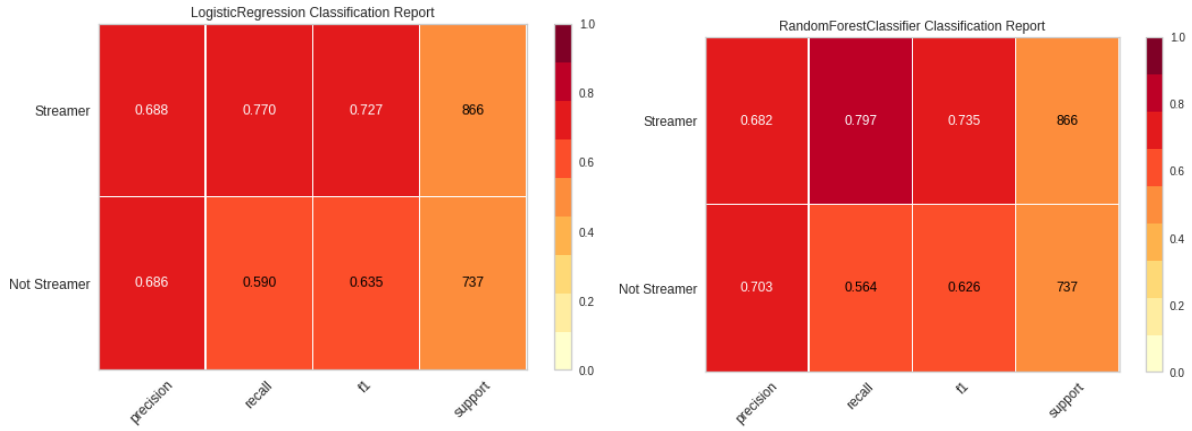
Results

As compared to the logistic regression model, the random forest classifier outperforms due to the higher AUC, F-1, and recall scores. Logistic regression outputs can be found in Appendix C.

Confusion Matrix

Random forest performs better than logistic regression in the confusion matrix as seen in Figure 11.

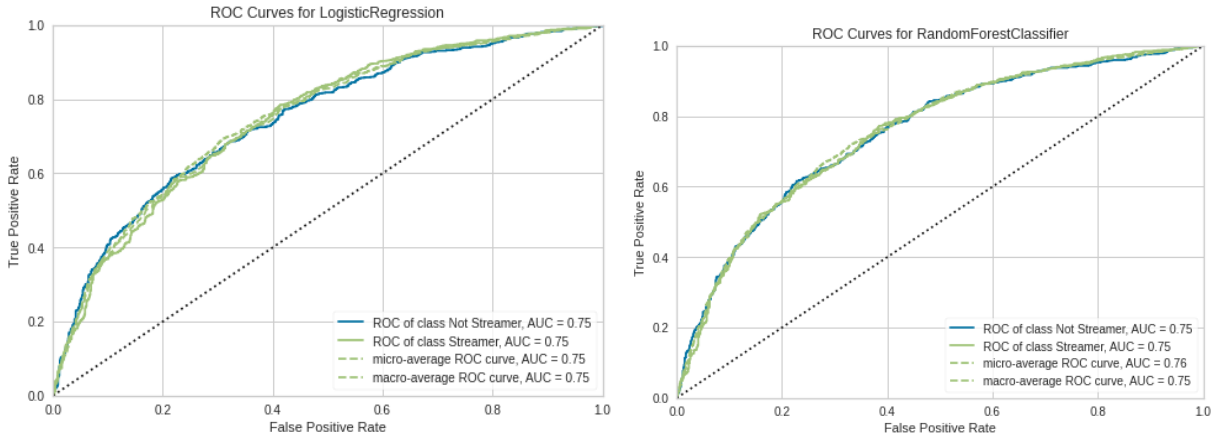
Figure 11 - Confusion Matrix, Logistic Regression vs. Random Forest



Receive Operator Curves (ROC)

We also see that AUC of the random forest was marginally higher than the AUC of logistic regression as seen in Figure 12.

Figure 12 - ROC, Logistic Regression vs. Random Forest

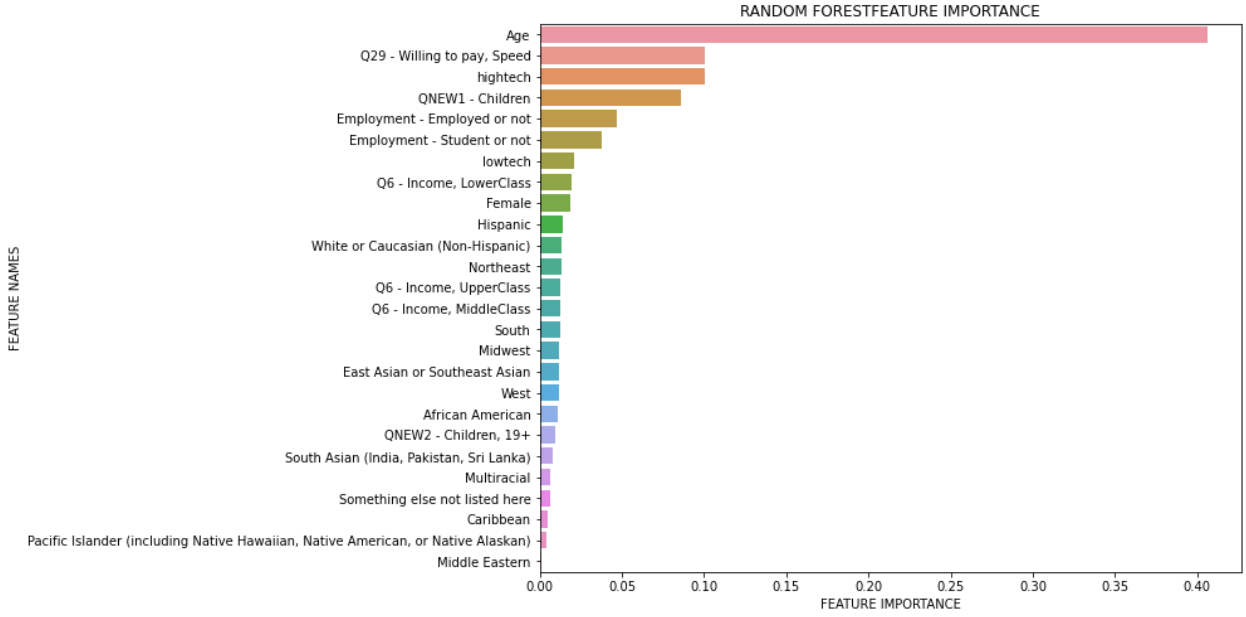


Both confusion matrix and ROC help us evaluate the classifiers and arrive at a decision that random forest is the way to go.

Feature Importance

In Figure 13, we see the rankings of features in order of importance produced by the best model.

Figure 13 - Feature Importance of Random Forest



Findings

Upon evaluating the best features in the random forest we are able to see consumers are likely to stream when:

- Regularly use high-tech equipment (i.e have drones, fitness watches, etc)

- Willing to pay for bandwidth (i.e. ready to pay for better Wifi Speeds)
- Aged below age 40 (i.e.are a younger audience)

Using these results and findings, we recommend specific courses of action for streaming companies.

Recommendations

Overall, the initial prediction model will help streaming companies uncover new consumers and advertise to them to increase consumer adoption, minimize the cost of acquisition, and help the company grow. earmark recommends streaming companies consider the following action items as they continue to carve out their niche in this competitive landscape.

Consumers predicted **"Does Not Have Streaming":**

- *Expand marketing efforts to untapped segments.* In particular, users aged 40-60 years old have the disposable income to pay for streaming services and are moving away from Pay TV. These users are an untapped segment that streaming companies can look for new consumers. They may be difficult to convert and require significant investment and research in UI/UX for this select age group.
- *Improving the value that streaming adds to consumers' life.* Increase the value that the streaming platform adds to the consumer's life. Currently, consumers think of streaming as a disposable service, but by adding more live news and educational content, the service can become more than entertainment and become an essential part of consumers' lives. We see this already occurring in the market with launches such as CNN+ and Peacock.
- *Partner with educational institutions.* Educational institutions are moving towards technology to transform and aid in the classroom experience. Like partnered banks, universities can offer preferred streaming services with select companies. At the same time, students are using streaming services but do not have individual accounts. The lack of personal accounts presents an opportunity for streaming companies to cross-sell through educational services.

Consumers predicted **"Has Streaming":**

- Advertise premium subscriptions to:
    - "High tech" individuals
    - Households with children less than 19 years old.
    - Individuals willing to pay for faster internet speed
- Incentivise Streaming to consumers of various income classes through customized Product Packaging:
    - A consumer's financial situation highly determines their choice to be a streaming subscriber

The streaming industry is becoming increasingly competitive, but utilizing machine learning models such as the ones outlined in this report will help differentiate streaming companies from their competition, decrease their cost of acquisition, and, increase their overall market share.

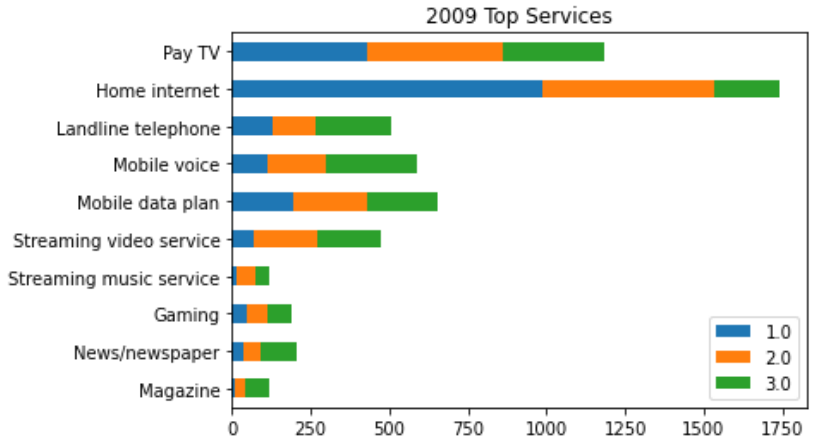
Division of Work

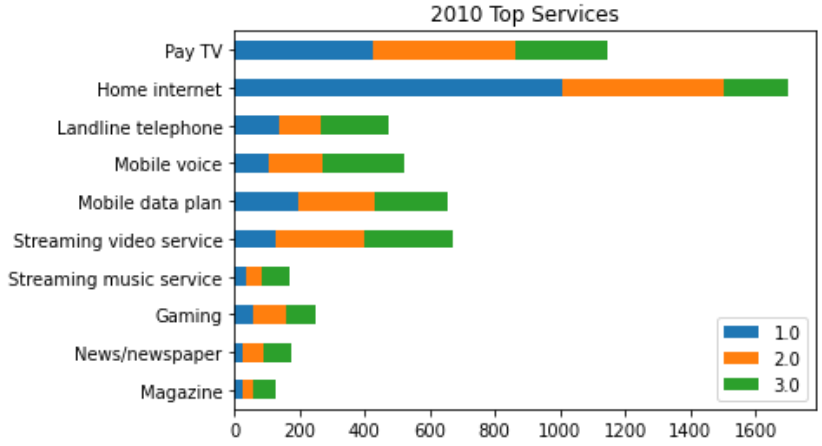
Each individual wrote the report and prepared slides for their respective parts.

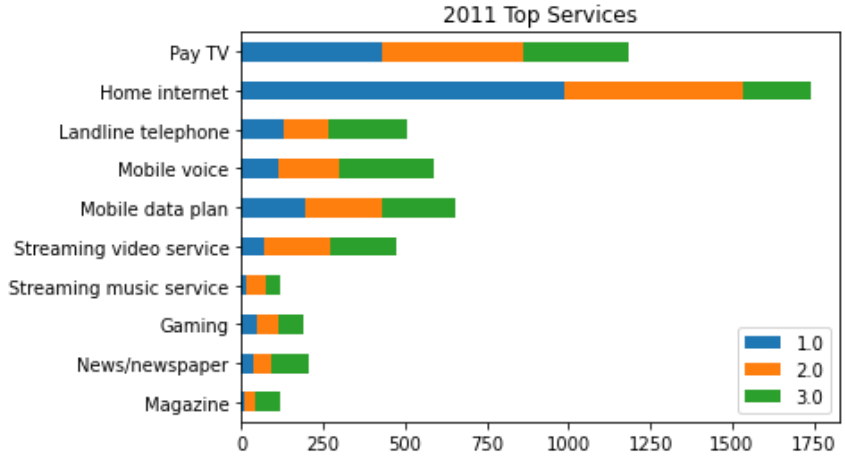
Task	**Responsible**
Business Case	Kelly
Exploratory Data Analysis	Cole & Kelly
Combining Datasets	Kelly
Feature Engineering	Cole & Michael
Outliers	Krish
Missing Values	Krish
Modeling	David
Evaluation	David
Recommendations	All

Appendix

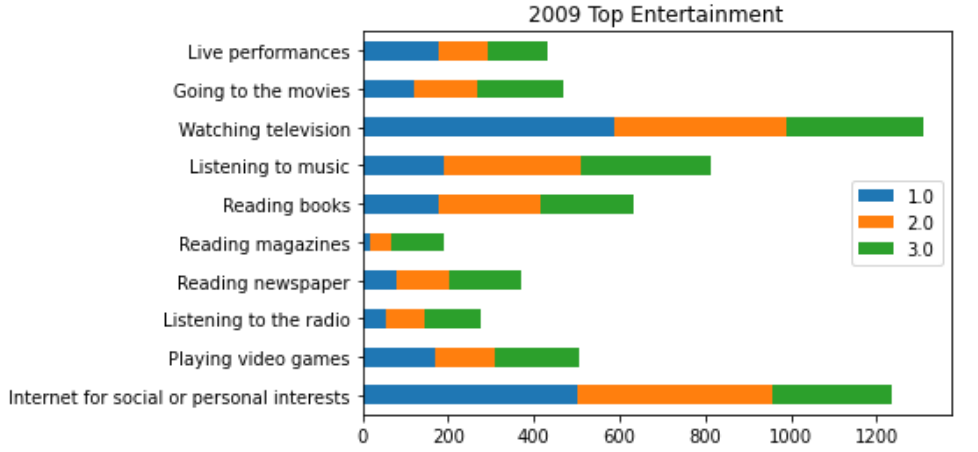
A - Top Valued Services

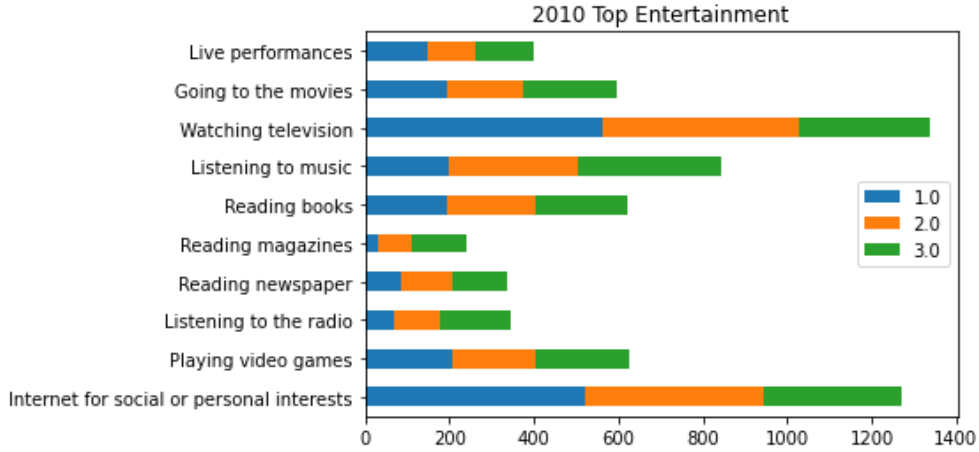


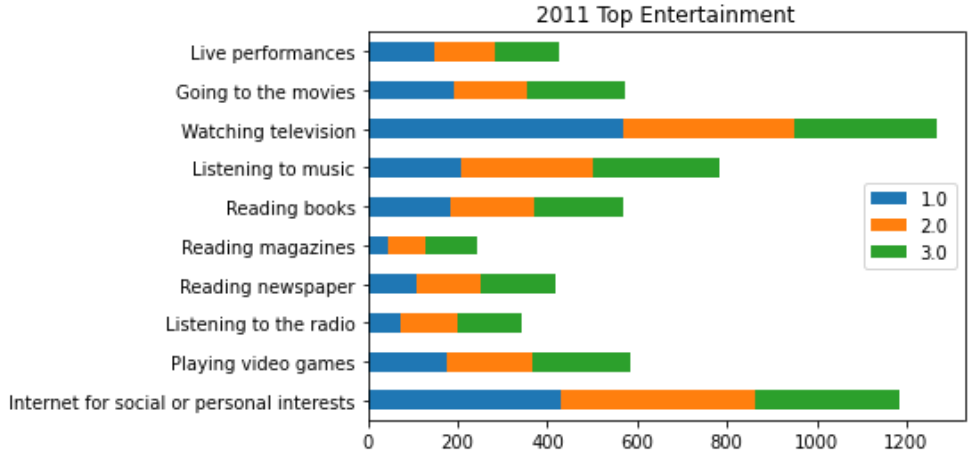




B - Top Entertainment

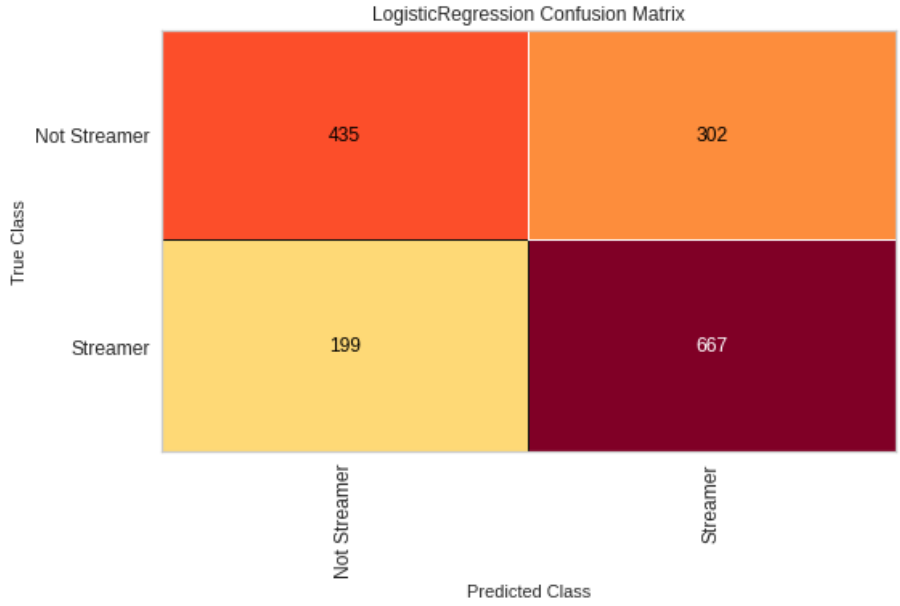




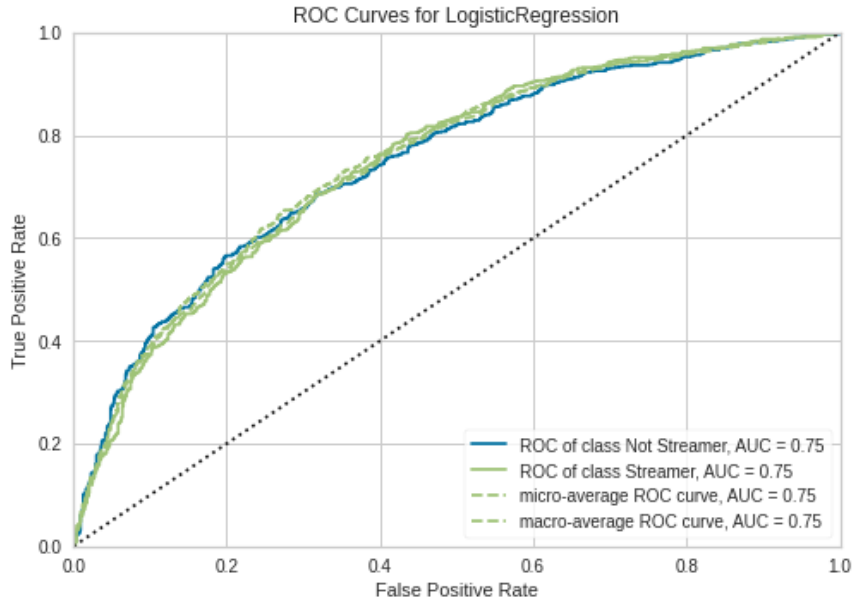


C - Grid Search, Logistic Regression Results

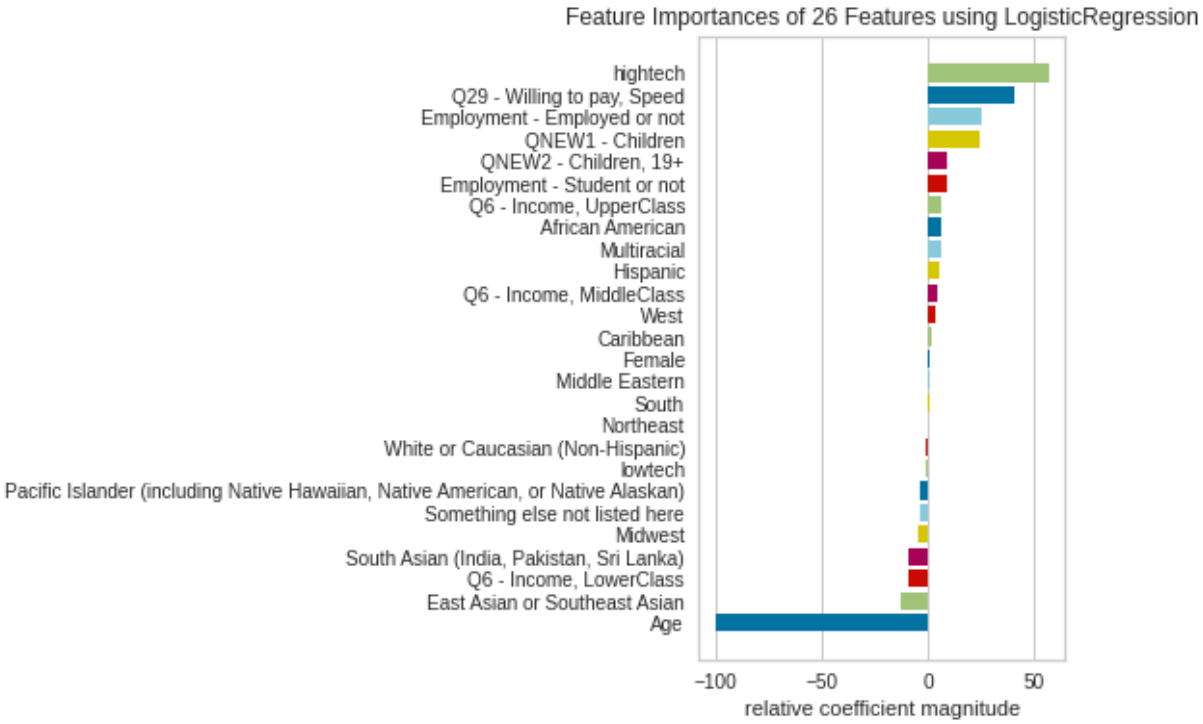
Confusion Matrix, Logistic Regression



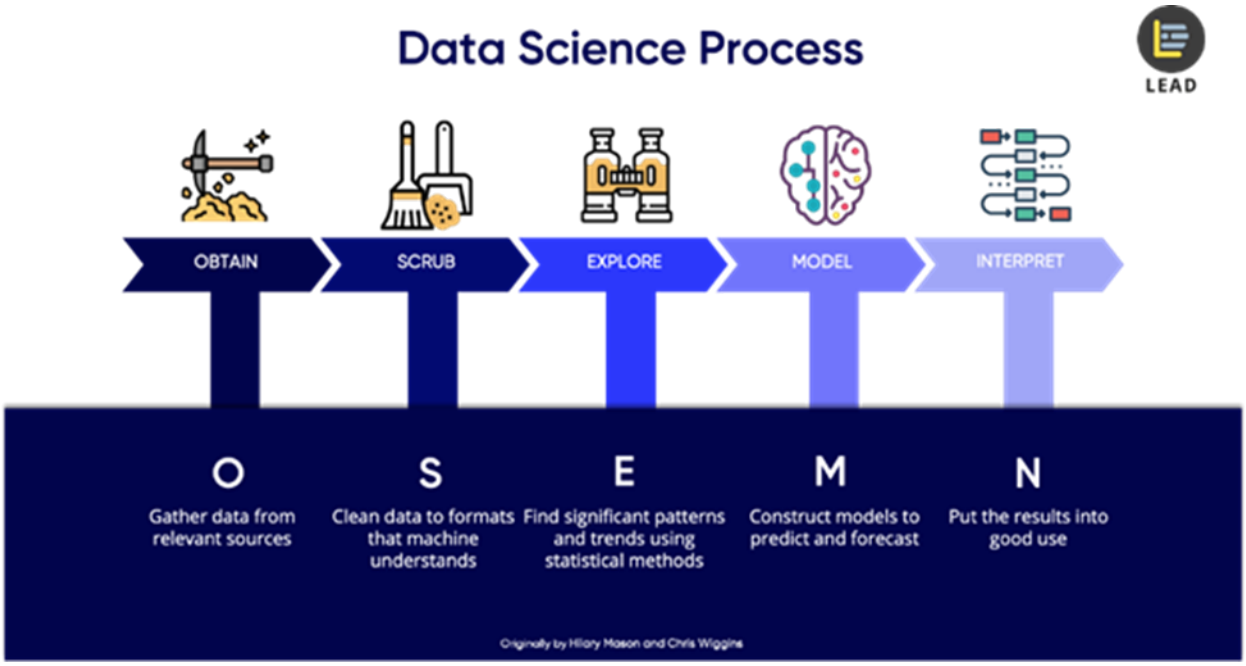
ROC Curve, Logistic Regression



Top Features, Logistic Regression

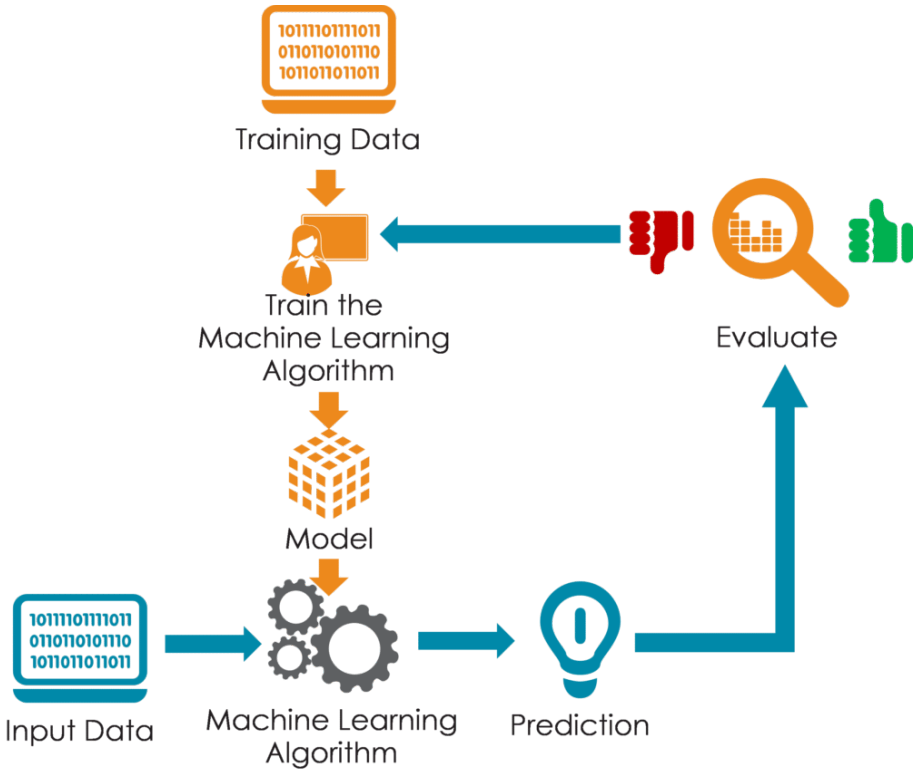


D - Data Science Process



E - Modelling Process

(Akella, 2022)



References

Akella, Bharani (2022, February). *Data Modeling in Data Science.* https://intellipaat.com/blog/tutorial/data-science-tutorial/modeling-the-data/

Cook, D. (2020, November). *Video Streaming Services in the US: US Specialized Industry Report OD6197.* IBISWorld. https://my-ibisworld-com.cmu.idm.oclc.org/us/en/industry-specialized/od6197/about

Deep AI. (n.d.) *Logistic Regression.* https://deepai.org/machine-learning-glossary-and-terms/logistic-regression

Lau, D. C. H. (2019, January 10). *5 steps of a data science project lifecycle*. Medium.]Retrieved February 28, 2022, from https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Lutins, Evan. (2017, September 5). *Grid Searching in Machine Learning: Quick Explanation and Python Implementation.* Medium. https://elutins.medium.com/grid-searching-in-machine-learning-quick-explanation-and-python-implementation-550552200596

Mitchell, Tom. (1997). *Machine Learning*. McGraw Hill. http://www.cs.cmu.edu/~tom/mlbook.html#:~:text=Machine%20Learning%2C%20Tom%20Mitchell%2C%20McGraw,that%20automatically%20learn%20users'%20interests.

Munagala, Ravali. (2021, March 19). Numpy Ninja. https://www.numpyninja.com/post/hyper-parameter-tuning-using-grid-search-and-random-search

Wood, Thomas. (n.d.) *Random Forests.* Deep AI. https://deepai.org/machine-learning-glossary-and-terms/random-forest