CAPSTONE PROJECT

Project name :- Android google play store applications analysis

Name :-Naresh Kumar

Student ID:- 0767830

Name of Supervisor:- Dr. Savita Seharawat

Date of Submission:- 31st January

Logo, company name

Description automatically generated with medium confidence

Table of Contents

[Abstract 3](#_Toc100775085)

[KEYWORDS: 3](#_Toc100775086)

[**RESEARCH QUESTIONS:** 3](#_Toc100775087)

[**TOOLS:** 4](#_Toc100775088)

[Methodology: 5](#_Toc100775089)

[Numerical Attributes: 11](#_Toc100775090)

[Table 1 Summary of numeric attributes:- 11](#_Toc100775091)

[Detailed Data Dictionary:- 11](#_Toc100775092)

[Exploratory Data Analysis:- 13](#_Toc100775093)

[FIGURE 1 . Maximum\_installs 13](#_Toc100775094)

[**FIGURE 2. Minimum\_installs** 14](#_Toc100775095)

[Data Visualization:- 14](#_Toc100775096)

[Train-Test-Split Method 18](#_Toc100775098)

[**Modelling** 18](#_Toc100775099)

[Stratified K Fold: 19](#_Toc100775100)

[Random Train-Test-Split 19](#_Toc100775101)

[Confusion Matrix Corresponding to Random Forest Classifier Algorithm:- 20](#_Toc100775102)

[FIGURE 7. Actual Values Vs Predicted Values 21](#_Toc100775103)

[Confusion Matrix Corresponding to Random Forest 21](#_Toc100775104)

[FIGURE 8. True Positive Rate Vs False Positive Rate 22](#_Toc100775105)

[Conclusion:- 22](#_Toc100775106)

[References:- 22](#_Toc100775107)

# Abstract

Our project name is Android Google Play applications analysis , moreover, known as the Google Play Store, and in which I used androiddataset.csv dataset in which 1048575 records and 24 attributes. With the help of android google play store we will download or purchase millions of apps, diversions, and other media onto your Android gadget. You'll be able discover programs for a wide cluster of interfaces. You'll be able download apps or recreations by exploring to the app or amusement page inside the Play Store and tapping Introduce. Numerous apps will be free, a few will have in-app promotions, a few will fetch cash, whereas others may offer in-app buys, or a combination of any of these things. For those who are interested, there's the Google Play Pass, which allows you to download hundreds of apps and recreations for complimentary, without advertisements or in-app buys.so in this data we have some important attributes and value that help us to understand all functions of this dataset. Its a stage. It’s Google’s stage for advertising different advanced substance to its customers. Opposite to what a few individuals may think, the Google Play Store isn't fair an app store, not at all. You'll discover all sorts of substance accessible here. The Google Play Store is domestic to music, motion pictures, books, and recreations

**KEYWORDS:** Data cleaning, data visualization, Logistic regression model, KNeighbors, Random Forest Classifier, Naïve bayes model , Sampling , K-Fold Cross Validation , Stratified K-Fold , Confusion Matrix

### **RESEARCH QUESTIONS:**

**Question 1)-** find how many applications are free, ad supported using pie chart?

**Solution** :- 99.4 % applications are free of cost for the customers , also same percentage is noticed in applications that supported ads that is 99.4% and 0.6% applications are paid and also not supported ads.

**Question – 2) Which category of content rating show highest numbers that shows rating corresponding to application.**

**Solution :- from all categories of content rating I find that people those are 18+ , they give more rating to the applications.**

**Question 3)** find accuracy of this project(dataset) by using different methods and also compare each other model to check which one is best as compare to the others?

**Solution :-** I used four models to find accuracy of this project That is Logistic regression , KNN model , Naïve bayes model and random forest classifier model .

Logistic regression accuracy is 59.33

KNN model accuracy is 54.91

Naïve bayes model accuracy is 59.09

Random forest classifier accuracy is 56.88

**So accuracy that we find with the help of logistic regression is best for our model that is 59.33**

### **TOOLS:**

We will use python software for performing operations and to visualize our dataset.

#### **GITHUB INFO:**

[**https://github.com/nareshsalhan141**](https://github.com/nareshsalhan141)

##### **LITERATURE REVIEW:**

From the research I found that **mr.Lavanya** define dataset as a name of google play store apps (**in 2019)** in which he was define lots of functions on the given dataset.in which he explain While numerous public datasets (on Kaggle and so forth) give Apple App Store information, there are very few partner datasets accessible for Google Play Store applications anyplace on the web. On burrowing further, I discovered that iTunes App Store page sends a well listed reference section like construction to consider basic and simple web scratching. Then again, Google Play Store utilizes complex advanced procedures (like unique page load) utilizing JQuery making scratching seriously testing.along with this google play store apps data cleaning dataset by **saba siddiqi** in which they explain about data cleaning **(in 2018).**they were explaining about questions and answer about given dataset that is also related to google play store apps .**In 2019 ,n1kshan** publish dataset related to playstore in which he define seaborn functions and also explain how show information via graphs , The google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an model. We have used a raw data set of Google Play Store from the Kaggle website. This data set contains 13 different features that can be used for predicting whether an app will be successful or not using different features. This data set is scraped from the Google Play Store. This journal talks about different classifier models that we used for prediction purposes and finding which one gives the highest accuracy. This journal also gives detailed information on feature extraction and the complete Data visualization done on this data set. [1] Kaggle.com.(2018). Google Play Store Apps.[online]https://www.kaggle.com/lava18/google-play-store-apps [Accessed 3 Mar. 2020].[2] “Mining and Analysis of Apps in Google Play,” Pro-ceedings of the 9th International Conference on WebInformation Systems and Technologies, 2013.[3] Google play store: number of apps2018(2018). [online] https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/ [Accessed 3 Mar. 2020].[4] Amit Chile, Dr. P. R. Gundalwar.(2019). Anal-ysis of Google Play Store Application.[online]http://ijraset.com/ﬁleserve.php?FID=24134 [Accessed3 Mar. 2020][5] T. Denoeux, M. Skarstein-Bjanger, Induction of deci-sion trees from partially classiﬁed data, in: Proceedingsof the 2000 IEEE International Conference on Systems,Man and Cybernetics (SMC’00), IEEE, Nashville, TN,2000, pp. 2923–2928.[6] Harman, M., Jia, Y., and Zhang, Y. (2012). App storemining and analysis: Msr for app stores. In 2012 9thIEEE Working Conference on Mining Software Repos-itories (MSR),pages 108–111.[7] R. P. Rajeswari, K. Juliet, and Aradhana, “TextClassiﬁcation for Student Data Set using NaiveBayes Classiﬁer and KNN Classiﬁer,” Int. J. Com-put. Trends Technol., vol. 43, no. 1, pp. 8–12, 2017.https://doi.org/10.14445/22312803/ijctt-v43p103[8] Jong, J. (2011). Predicting rating with sentiment anal-ysis. [online] http://cs229.stanford. edu/proj2011/Jong-PredictingRatingwithSentimentAnalysis.pdf.[9] [2015].Grover, S. 3 apps that failed (and whatthey teach us about app marketing). [online]https://blog.placeit.net/apps-fail-teach-us-app-marketing

## **Methodology:**

Firstly, our dataset is uncleaned like there are some missing values . First, we will clean our dataset to perform father operations. In methodology part we will use regression model to find out relationship between dependent and independent variables.

**DATA PROCESSING**

**EXPERIMENTAL DESIGN**

**MODELING IMPLEMENTATION AND VALUES**

**CONCLUSION**

**DETAIL DATA :-DATA DICTIONARY HERE :-**

In detailed data dictionary, we check the different levels of categories of an attribute and also total

number of counts in each. We also check the datatype of each attribute and convert them into

appropriate type to get desired results. So, we convert object attributes into category and float into

integer in order to get accurate result.

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes name** | **description** | **datatype** | **Count** |
|  |  |  | (Label name=count value) |
| App name | Define name of each application like call of duty.pub-g etc | object | Flashlight = 23  Calculator = 17  Gallery = 11 |
| App id | It define ID of the application that is unique for every application | object | com.ishakwe.gakondo = 1  petrosidagresik.sidacare =1  com.appsislamik.quranne =1 |
| category | It shows categories of application that stored in this dataset. | object | Education = 14751  Music & Audio = 9801  Tools = 9012  Business = 8963 |
| Rating | It shows rating of each application like 4.4, 7,9 etc. | Float64 | 0.0 = 65845  5.0 = 6146  4.2 = 5469  4.4 = 5375 |
|  |  |  |
|  |  |  |
|  |  |  |
| Rating count | It defines how many times user give ratings to each application. | Float64 | 0.0 = 65845  5.0 = 3991  6.0 = 3328  7.0 = 2940 |
|  |  |  |
|  |  |  |
|  |  |  |
| installs | How many times people install each application |  | 100+ = 27559  1,000+ = 24737  10+ = 18695  10,000+ = 16048 |
|  |  | object |
|  |  |  |
|  |  |  |
| Minimum installs | Minimum no of install apps | Float64 | 1.000000e+02 = 27559  1.000000e+03 = 24737  1.000000e+01 = 18695  1.000000e+04 = 16048 |
|  |  |  |
|  |  |  |
|  |  |  |
| Maximum installs | Shows no of apps that maximum install. | Float64 | 5.0 = 1062  2.0 = 1040  3.0 = 1037  4.0 = 1023 |
|  |  |  |
|  |  |  |
|  |  |  |
| Free | It shows application is free or not. | category | True = 140745  False = 2863 |
|  |  |  |
| Price | It define price of each application. | Float64 | 0.000000 = 140755  0.990000 = 786  1.990000 = 388  2.990000 = 253 |
|  |  |  |
|  |  |  |
|  |  |  |
| Currency | It shows currency type that used to buy application. | category | USD = 142998  XXX = 592  INR = 2  BRL = 1 |
|  |  |  |
|  |  |  |
|  |  |  |
| Size | Shows size of application. | category | Varies with device = 4538  11M = 3826  12M = 3433  13M = 2972 |
|  |  |  |
|  |  |  |
|  |  |  |
| Minimum android(version) | Shows version of application like 4.4 , 5,5. | object | 4.1 and up = 37546  5.0 and up = 24591  4.4 and up = 24183  4.0.3 and up = 11227 |
|  |  |  |
|  |  |  |
|  |  |  |
| Developer id | Define developer id that published their app in play store. | object | TRAINERIZE = 339 |
|  |  |  | Subsplash Inc = 338 |
|  |  |  | ChowNow = 310 |
|  |  |  | Phorest = 176 |
| Developer websites | It shows website of each developer that makes apps. | object | [http://www.subsplash.com/ = 454](http://www.subsplash.com/) |
|  |  |  | [http://www.chownow.com = 286](http://www.chownow.com/) |
|  |  |  | [http://www.sharefaith.com/category/church-websites.html = 119](http://www.sharefaith.com/category/church-websites.html) |
|  |  |  | [https://foodsoul.pro = 113](https://foodsoul.pro/) |
| Developer email | Define email id of developer. | object | [support@classplus.co = 639](mailto:support@classplus.co) |
|  |  |  | [appsupport@subsplash.com = 458](mailto:appsupport@subsplash.com) |
|  |  |  | [help@trainerize.com = 315](mailto:help@trainerize.com) |
|  |  |  | [eng-android@chownow.com = 310](mailto:eng-android@chownow.com) |
| Released | Define release date of application. | category | 15-Jun-20 = 145 |
|  |  |  | 18-Feb-20 = 138 |
|  |  |  | 9-Jun-20 = 138 |
|  |  |  | 5-Jun-20 = 138 |
| Last updated | Define date when application is updated | object | 11-Jun-21 = 666 |
|  |  |  | 9-Jun-21 = 661 |
|  |  |  | 10-Jun-21 = 644 |
|  |  |  | 14-Jun-21 = 627 |
| Content rating | Define age group that gives rating for applications. | object | Everyone = 125710 |
|  |  |  | Teen = 12158 |
|  |  |  | Mature 17+ = 3657 |
|  |  |  | Everyone 10+ = 2065 |
| Privacy policy | Shows instruction of privacy policy to the user. | object | [http://www.subsplash.com/legal/privacy 447](http://www.subsplash.com/legal/privacy) |
|  |  |  | [http://www.trainerize.com/privacy.aspx 340](http://www.trainerize.com/privacy.aspx) |
|  |  |  | [https://unity3d.com/legal/privacy-policy 338](https://unity3d.com/legal/privacy-policy) |
|  |  |  | [http://www.chownow.com/privacy-policy 304](http://www.chownow.com/privacy-policy) |
| Ad supported | Shows app is ad supported or not | object | False = 72010 |
|  |  |  | True = 71598 |
| In app purchases | Define application in purchases or not. | object | False = 131555 |
|  |  |  | True = 12053 |
| Editor’s choice | It shows application under editor choice or not. | object | False = 143563 |
|  |  |  | True = 45 |
| Scraped Time | It shows scraped time of application. | object | 6/15/2021 21:59 = 1107 |
|  |  |  | 6/15/2021 22:33 = 1095 |
|  |  |  | 6/15/2021 22:17 = 1095 |
|  |  |  | 6/15/2021 21:42 = 1095 |

## **Numerical Attributes**:

## Table 1 **Summary of numeric attributes:-**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Attributes Name | count | mean | std | min | 25% | 50% | 75% | max |
| Rating | 1038265 | 2 | 2 | 0 | 0 | 3 | 4 | 5 |
| Rating\_Count | 1038265 | 2937 | 255967 | 0 | 0 | 6 | 42 | 1.39E+08 |
| Minimum\_Installs | 1048529 | 176200 | 13259642 | 0 | 50 | 500 | 5000 | 5E+09 |
| Maximum\_Installs | 1048575 | 304457 | 20466334 | 0 | 84 | 695 | 7361 | 9.77E+09 |
| Price | 1048575 | 0 | 3 | 0 | 0 | 0 | 0 | 400 |

contain summary of dataset it is showing the numeric attributes of count, mean or average, is calculated by adding up the scores and dividing the total by the number of scores, standard deviation is a measure of dispersion of data values from the mean.

## **Detailed Data Dictionary:-**

In this step, I took one variable at a time and do the changes in each attributes I change the datatype corresponding to my dataset. For numeric attribute, I created five number of summary and find the count, mean. Standard deviation of each attribute. After that I checked the quantity of outliers in each numeric attribute. Next, I checked levels, counts and categories of columns and attribute along with imbalance in the dataset.

**Missing data:-**

Missing values are the null values in a dataset. When there is no value stored for any variable in an observation then that condition is treated as null/missing or N/A value. If there is an availability of null values in any dataset, then our result will never be accurate. we can see that our dataset contains sufficient amount of NA values and to perform analysis we need to remove these missing values from dataset by using suitable criteria. If we are not removing these null values, then we get inaccurate output.

|  |  |
| --- | --- |
| **Developer\_Website** | 100692 |
| **Privacy\_Policy** | 55600 |
| **Released** | 9349 |
| **Rating** | 3033 |
| **Rating\_Count** | 3033 |

Figure that showing missing values

**Criteria for cleaning the missing values:**

Next step is to analyze if there is any pattern for the missing values then I remove it for further analysis.I drop those attributes with machine algorithm techniques those have sufficient number of NA values. After that I replacing NA values in each attribute. Criteria for cleaning the missing values:As we can see ‘Developer\_website’, ‘Privacy\_policy’, ‘Released’, ‘Rating’, ‘Rating\_Count’ contains more than 10000 null values so if we randomly drop N/A values without any argument, it will remove the whole row wherever missing value exist, then almost half of our data will be deleted. That is why for that we dropped those attributes which contains highest number (10000 or more) of missing values in our dataset and less than 10000 missing values, we dropped the whole row where NA value exists.

# **Exploratory Data Analysis:-**

Chart, box and whisker chart

Description automatically generated

## **FIGURE 1 . Maximum\_installs**

Above boxplot shows positively skewed distribution as it shows that median is more closer to the bottom of the box and lower values towards lower bound. It represent most of the values occurred between 0 to 1000.It also contains huge amount of outliers at the upper bound.

Chart, scatter chart

Description automatically generated

### **FIGURE 2. Minimum\_installs**

The above plot is positive skewed because median is more closer to lower bound. Its most of the values occurred between 0 to 1000 and also contains few outliers towards upper bound.

# **Data Visualization:-**

**Counts vs currency:**

Chart, bar chart

Description automatically generated

#### **FIGURE 3 . Count vs currency**

This bar graph represents the information between currency and counts. We can see that from this graph that the counts of USD currency has highest as compared to others. It is nearly 700000.

**Show highest Content rating category :-**

Chart, bar chart

Description automatically generated

#### **FIGURE 4. Count vs Content\_Rating**

We can noticed that people who are 18+ that people give highest rating to the content of all application as compare to the other categories( adults only , everyone , mature , 17+ , Teen , Unrated) like more then as a 60,00,00 counts number

**Show how many applications are supported purchased feature**.:-

\Chart, pie chart

Description automatically generated

**FIGURE 5. Is App Supported or Not?**

In above pie chart the blue colour represent the percentage of apps which are purchased that is 99.4% and the yellow colour represents the apps which are not purchased that are 0.6%.

**Show how many applications are free of cost to the customer:-**

\]Chart, pie chart

Description automatically generated

##### **FIGURE 6. Is App Free or Not?**

The above pie chart is showing the percentage of apps that are free or not free. There are 99.4% apps that are free of cost. And the rest are not free.

## **Modeling Implementation**:-

## **Train-Test-Split Method**

In train-test-split method the entire dataset is partitioned into training and testing sets. The training set contain 70% records and, in the testing, set it contain 30% records. Then I apply the smote technique to make my target variable balance.

### **Modelling**

In the modelling part there are four methods Logistic Regression, KNN, Naive Bayes, random forest that I have applied on my dataset.

**Logistic Regression**:

This model is supervised learning classification algorithm. It is used to predict the value of target variable.

**KNN**:

The KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

**Naive Bayes**:

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.

**Random Forest CLASSIFIER**:

It is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

**Cross Validation:-** It is a resampling method that uses different portions of the data to test and train a model on different iterations.

## **Stratified K Fold:**

This pass-validation object returns stratified folds and is a variation of K-Fold. The folds are made by using preserving the share of samples in every class. I divided the data into 5 stratified folds. The 4 folds are then used to healthy the version, and the 5th fold is used to check it. Repeat until each fold has been used as a take a look at set. Then add collectively all of the effects and calculate the common. That might be the version's metric of success.

## **Random Train-Test-Split**

This method combines the k-fold-cross-validation approach with typical train-test-splits. I create random divides of the information in the training-check set, similar to the move-validation technique, and then repeat the procedure of splitting and trying out the algorithm many times. I divided the statistics into five Repeated Random Test-Train Splits.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | TRAIN\_TEST\_SPLIT | KFOLDS\_5 | STRATIFIEDKFOLD\_5 | RRTESTTRAINSPLITS\_5 |
| Logistic Regression | 59.22 | 52.97 | 52.88 | 53.69 |
| KNN | 53.83 | 51.02 | 56.03 | 58.80 |
| Naïve Bayes | 58.92 | 53.88 | 54.38 | 58.54 |
| Random Forest | 57.43 | 53.50 | 56.07 | 58.12 |

#### **Table 2 Accuracy Matrix**

## **Confusion Matrix Corresponding to Random Forest Classifier Algorithm:-**

Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.and it a tabular summary of the number of correct and incorrect predictions made by a classifier.

Chart, treemap chart

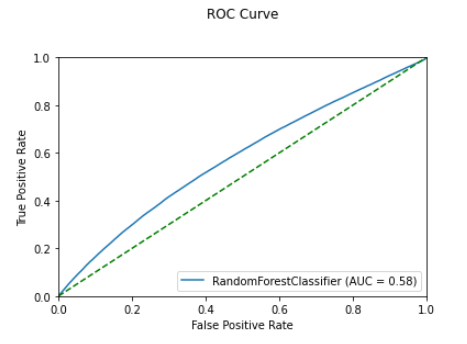
Description automatically generated

## **FIGURE 7. Actual Values Vs Predicted Values**

## **Confusion Matrix Corresponding to Random Forest**

When the target variable's actual value is 0 and the predicted value is also 0 and 38.60% of observation fall into first quadrant. 18.20% of observations fall into the second quadrant; whilst the target variable's actual value is 0 and predicted value is 1. 24.37% of observations fall into the third quadrant; and while the target variable's actual value is 1 and predicted value 0, 18.83% of observations fall into the fourth quadrant. The actual and predicted values are each 1.

**ROC CURVE (RECEIVER OPERATING CHARACTERISTIC) :-**Roc curves are widely used in binary classification to study the output of a classifier



## **FIGURE 8. True Positive Rate Vs False Positive Rate**

FROM ABOVE GRAPH, AREA UNDER THE CURVE SCORE IS 0.58. IT SHOW THAT IT IS AN OPTIMAL CLASSIFIER.

# **Conclusion:-**

In conclusion I can say that i perform all possible operations and algorithms to find some basic things from our project and after finding da accuracy of this project by using four models , I find good result that means its easy to use in the future. Secondly By using pie charts and bar graphs i defined visualization of some few attributes that help us to see direct results like minimum and maximum values from the different categories.

## **References:-**

[*"Google Play Store APK"*](https://www.apkmirror.com/apk/google-inc/google-play-store/google-play-store-29-0-14-release/google-play-store-29-0-14-21-0-pr-423376132-android-apk-download/).

Jump up to[*"Info about the Google Play Music phase-out"*](https://support.google.com/googleplaymusic/answer/9973710?hl=en-GB&ref_topic=6230811). Google Help*. Retrieved November 13, 2020*.

[*"Number of Google Play Store apps 2017 | Statistic"*](https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/). Statista*. Retrieved January 3, 2018*.

[*"Number of Android applications on the Google Play store"*](https://www.appbrain.com/stats/number-of-android-apps). appbrain.com*. Retrieved August 12, 2020*.

Jump up to[*"Number of Android applications"*](https://web.archive.org/web/20170210051327/https:/www.appbrain.com/stats/number-of-android-apps). AppBrain. February 9, 2017. Archived from [*the original*](https://www.appbrain.com/stats/number-of-android-apps) on February 10, 2017*. Retrieved February 24, 2017*.

 Kastrenakes, Jacob (February 28, 2017). [*"Google now lets apps display sale prices in the Play Store"*](https://www.theverge.com/2017/2/28/14760940/google-play-store-app-sales). [*The Verge*](https://en.wikipedia.org/wiki/The_Verge). [*Vox Media*](https://en.wikipedia.org/wiki/Vox_Media)*. Retrieved March 1, 2017*.

 Rossignol, Derrick (February 28, 2017). [*"Google now lets developers offer sales on Android apps"*](https://www.engadget.com/2017/02/28/google-now-lets-developers-offer-sales-on-android-apps/). *[Engadget](https://en.wikipedia.org/wiki/Engadget" \o "Engadget)*. [*AOL*](https://en.wikipedia.org/wiki/AOL)*. Retrieved March 1, 2017*.

[*"How to use a Google Play gift card"*](https://www.androidcentral.com/how-use-google-play-gift-card). Android Central. February 5, 2019*. Retrieved June 30, 2020*.

O'Brien, Terrence (May 2, 2012). [*"Google Play adds carrier billing for music, movies and books"*](https://www.engadget.com/2012/05/02/google-play-adds-carrier-billing-for-music-movies-and-books/). *[Engadget](https://en.wikipedia.org/wiki/Engadget" \o "Engadget)*. [*AOL*](https://en.wikipedia.org/wiki/AOL)*. Retrieved February 23, 2017*.

[*"Two Years Later, Google Play Support Docs Acknowledge 48-Hour Refund Window"*](https://www.androidpolice.com/2016/06/02/two-years-later-google-play-support-docs-acknowledge-48-hour-refund-window/). Android Police. June 2, 2016*. Retrieved June 30, 2020*.