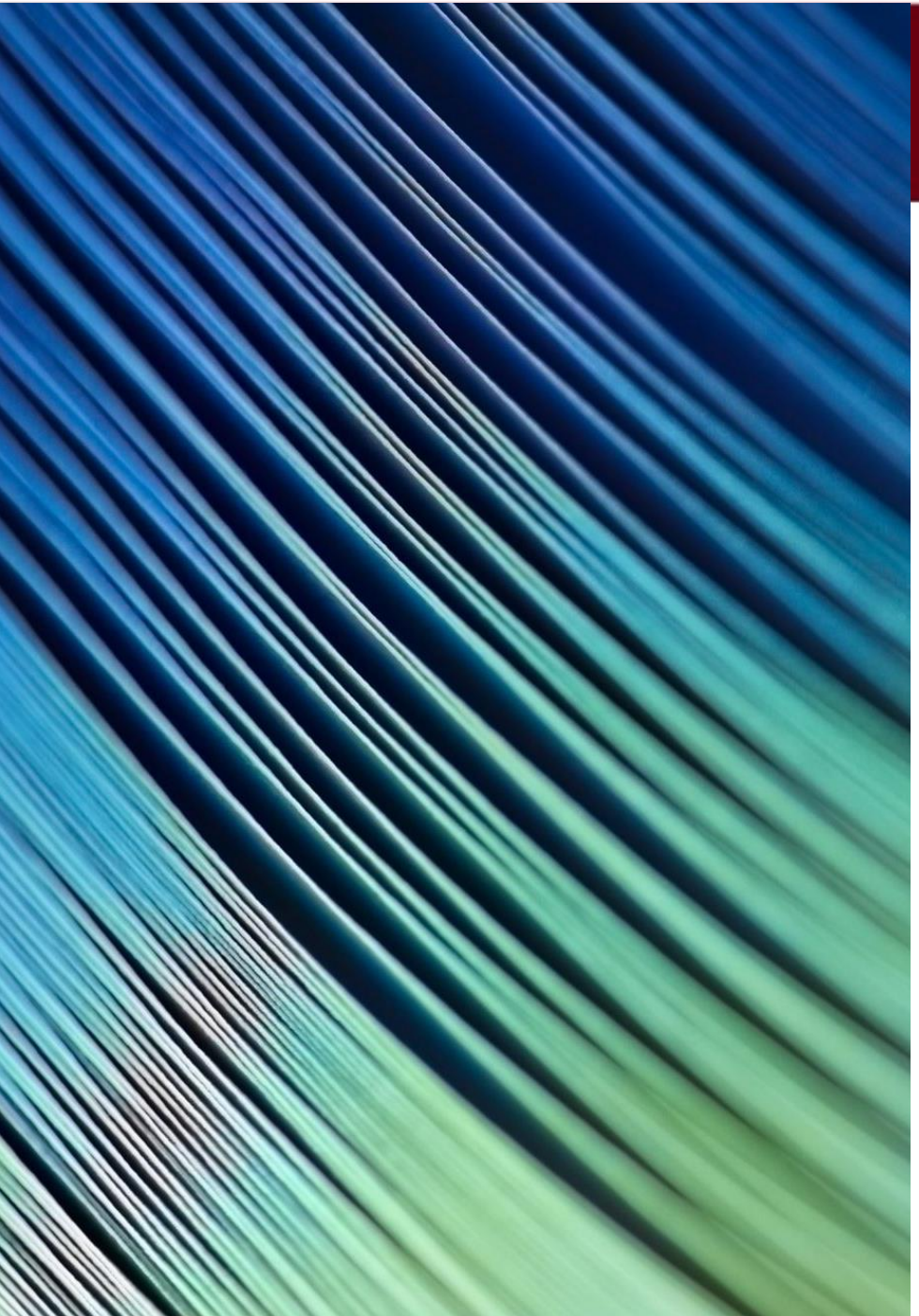
Predictions of App Installs in Play Store

Team 11

Vincentia Angelica
Cloris He
Martin Gui
Poonam S. Ahuja



Agenda

Executive Summary

Data preprocessing/ processing

Data modeling

Data findings

Conclusion & way ahead

Executive Summary

Main goal is to aid new playstore app success by predicting number of app installs

Background

- Playstore apps - enormous potential to drive mobile app business to success
- Immense competition, important for developers to channelize energy towards success

Goal

- Help in predicting the range of installs for a new app
- Better assist the developers to alter the feature mix for success

Proposal

- Build machine learning models to aid developers for a successful app by predicting app installs
- Data exploration, isolation of relevant features for better prediction

Why predicting range of 'Installs'?

BLUECLOUD

How Many Downloads Should My App Get?

PUBLISHING APPS & DOWNLOADS

So you've built your app and everything's ready to go. You glance over the calendar and notice there's a 3-day weekend starting tomorrow. Perfect! You click "Publish" and call it a day.

The next day you open your AppAnnie Report only to see 46 downloads. What!?! How could this happen? I spent $4,000 on this app, it's a beast! I've talked to dozens of top developers and read countless blogs. I even purchased this source code from a reputable developer with a proven track record.

Ok, chill Mr. Psychotic Developer. Lets break this down…

If you've released an app before, you know the first 3-4 days of your release are weighted the most. App releases typically work like a bell curve. **It's common to receive a really low download count this first day.** This could be because of various ASO reasons, Apple processing your app, or even the time of day it was published. Don't trip.

The next day or two you typically see a higher download number. It could be X2 or maybe X50, but **it's almost always higher.** This is Apple computing all your data and testing you out.

Day 3-5 typically marks the end of your app's push from Apple. These are typically the realistic download numbers you can expect to receive moving forward. If you're back at 46 downloads, it doesnt necessarily mean your app is dead. Hell, if it's a paid app at $2.99, you're profiting $138 a day.

Data summary

Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	7-Jan-18	1.0.0	4.0.3 and up
AUTO_AND_VEHICLES	3.9	967	14M	500,000+	Paid	$4.99	Teen	Art & Design;Pretend Play	15-Jan-18	2.0.0	4.2 and up
BEAUTY	4.7	87510	8.7M	5,000,000+	NaN	$3.99	Everyone 10+	Art & Design;Creativity	1-Aug-18	1.2.4	4.4 and up
BOOKS_AND_REFERENCE	4.5	215644	25M	50,000,000+	0	$6.99	Mature 17+	Art & Design;Action & Adventure	8-Jun-18	Varies with d	2.3 and up
BUSINESS	4.3	167	2.8M	100,000+		$1.49	Adults only 18+	Auto & Vehicles	20-Jun-18	1.1	3.0 and up
COMICS	4.4	178	5.6M	50,000+		$2.99	Unrated	Beauty	26-Mar-17	1	4.1 and up
COMMUNICATION	3.8	36815	29M	1,000,000+		$7.99		Books & Reference	26-Apr-18	6.1.61.1	4.0 and up
DATING	4.2	13791	33M	10,000,000+		$5.99		Business	14-Jun-18	2.9.2	2.3.3 and up
EDUCATION	4.6	121	3.1M	5,000+		$3.49		Comics	20-Sep-17	2.8	Varies with de
ENTERTAINMENT	3.2	13880	28M	100,000,000+		$1.99		Comics;Creativity	3-Jul-18	1.0.4	2.2 and up
EVENTS	4	8788	12M	1,000,000,000+		$9.99		Communication	27-Oct-17	1.0.15	5.0 and up
FINANCE	NaN	44829	20M	1,000+		$7.49		Dating	31-Jul-18	3.8	6.0 and up
FOOD_AND_DRINK	4.8	4326	21M	500,000,000+		$0.99		Education;Education	2-Apr-18	1.2.3	1.6 and up
HEALTH_AND_FITNESS	4.9	1518	37M	50+		$9.00		Education	26-Jun-18	NaN	1.5 and up
HOUSE_AND_HOME	3.6	55	2.7M	100+		$5.49		Education;Creativity	3-Aug-18	3.1	2.1 and up
LIBRARIES_AND_DEMO	3.7	3632	5.5M	500+		$10.00		Education;Music & Video	6-Jun-18	2.2.5	7.0 and up
LIFESTYLE	3.3	27	17M	10+		$24.99		Education;Action & Adventure	7-Nov-17	5.5.4	5.1 and up
GAME	3.4	194216	39M	1+		$11.99		Education;Pretend Play	30-Jul-18	4	4.3 and up
FAMILY	3.5	224399	31M	5+		$79.99		Education;Brain Games	20-Apr-18	2.2.6.2	4.0.3 - 7.1.1
MEDICAL	3.1	450	4.2M	0+		$16.99		Entertainment	20-Mar-18	1.1.3	2.0 and up
SOCIAL	5	654	7.0M	0		$14.99		Entertainment;Music & Video	12-Jul-18	1.5	3.2 and up
SHOPPING	2.6	7699	23M			$1.00		Entertainment;Brain Games	7-Mar-18	1.0.8	4.4W and up
PHOTOGRAPHY	3	61	6.0M			$29.99		Entertainment;Creativity	7-Jul-18	1.03	7.1 and up
SPORTS	1.9	118	6.1M			$12.99		Events	25-Apr-18	6	7.0 - 7.1.1

Class

Attributes

Categorical: Category, Type, Content Rating, Genres(have multiple items in one cell), Installs

Continuous: Rating, Reviews, Size, Price, Last Updated , Current Ver, Android Ver

Algorithms - Naïve Bayes vs Random Forest

Naïve Bayes

+ Good with relatively small features
+ Good with categorical data
+ Fast
- Features assumed to be independent
- It gives all features the same level of importance

Random Forest

+ Use multiple decision tree to have a better accuracy
+ Good for classification
+ Works well with both categorical /numerical
+ Reduce overfitting (compare to decision tree)
- However, if the model is not carefully built, it will have an overfitting problem
- The algorithm computation is complex, so it's slower

Steps

Run Naïve Bayes and Random Forest algorithm preprocessing to make predictions

Data Cleaning

Run Naïve Bayes and Random Forest post processing to make predictions

Data pre-processing (1/6)

Naïve Bayes

Correctly Classified Instances	1930	17.8028 %
Incorrectly Classified Instances	8911	82.1972 %

› Random Forest:

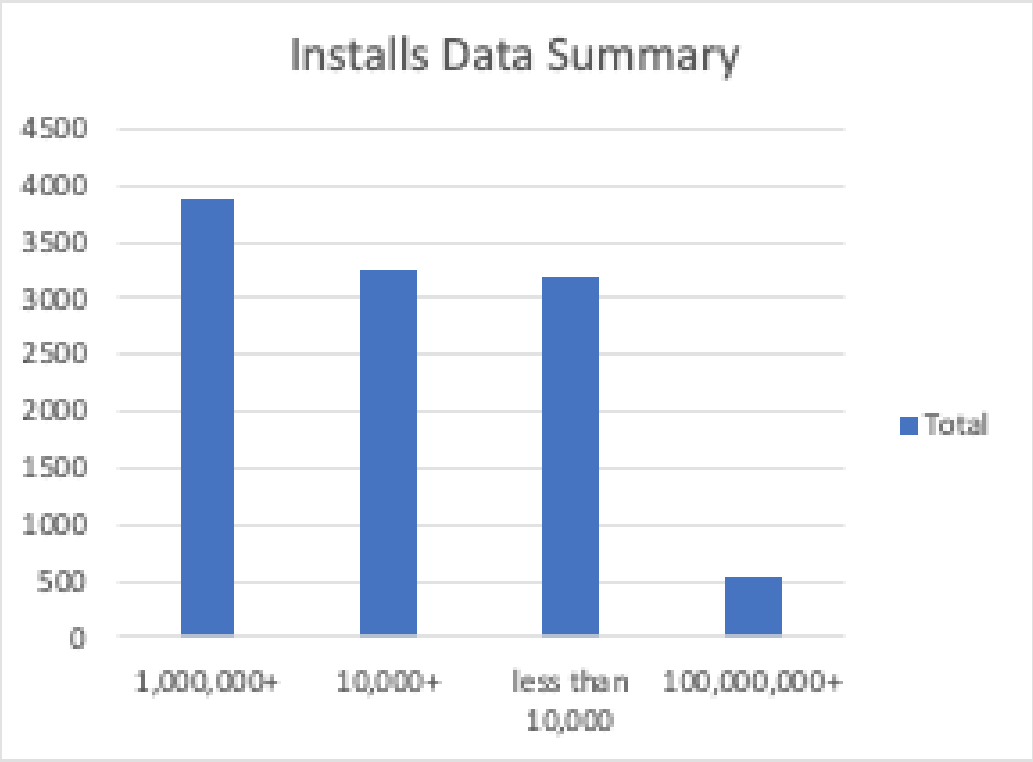
Correctly Classified Instances	2486	22.9315 %
Incorrectly Classified Instances	8355	77.0685 %

Data pre-processing (2/6)

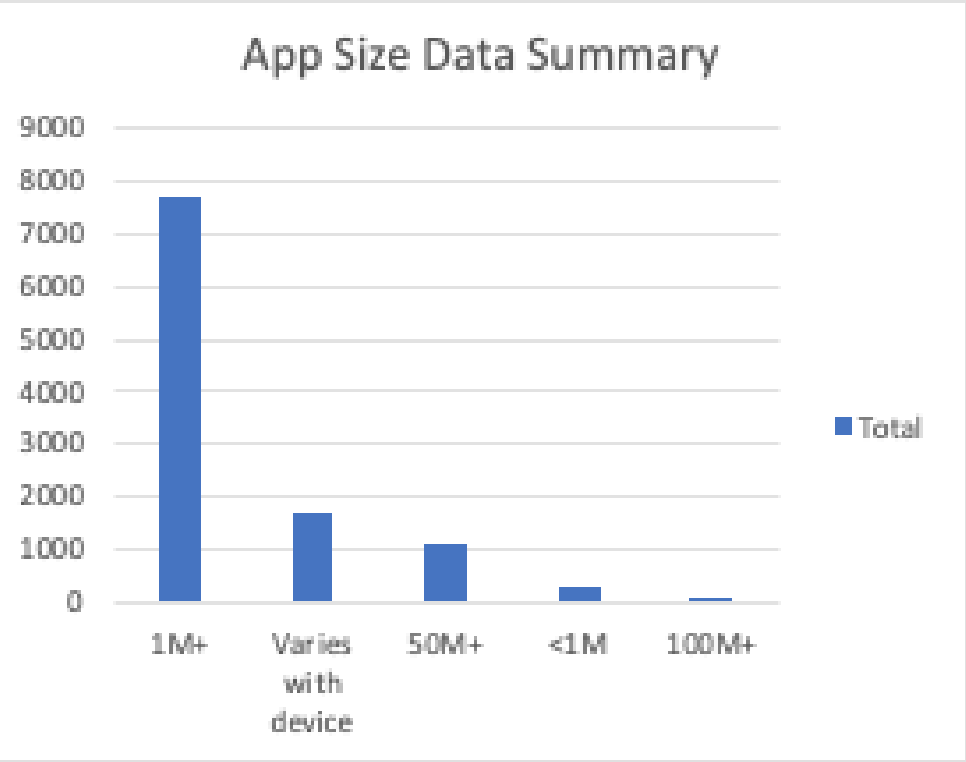
› **Bin Size + Installs**

Installs
10,000+
500,000+
5,000,000+
50,000,000+
100,000+
50,000+
1,000,000+
10,000,000+
5,000+
100,000,000+
1,000,000,000+
1,000+
500,000,000+
50+
100+
500+
10+
1+
5+
0+
0

21 class --> 4 class



Continuous -> Categorical



Data pre-processing (3/6)

- **Bin Size + Installs**
- Naïve Bayes

```
Correctly Classified Instances        4665               43.0311 %
Incorrectly Classified Instances      6176               56.9689 %
```

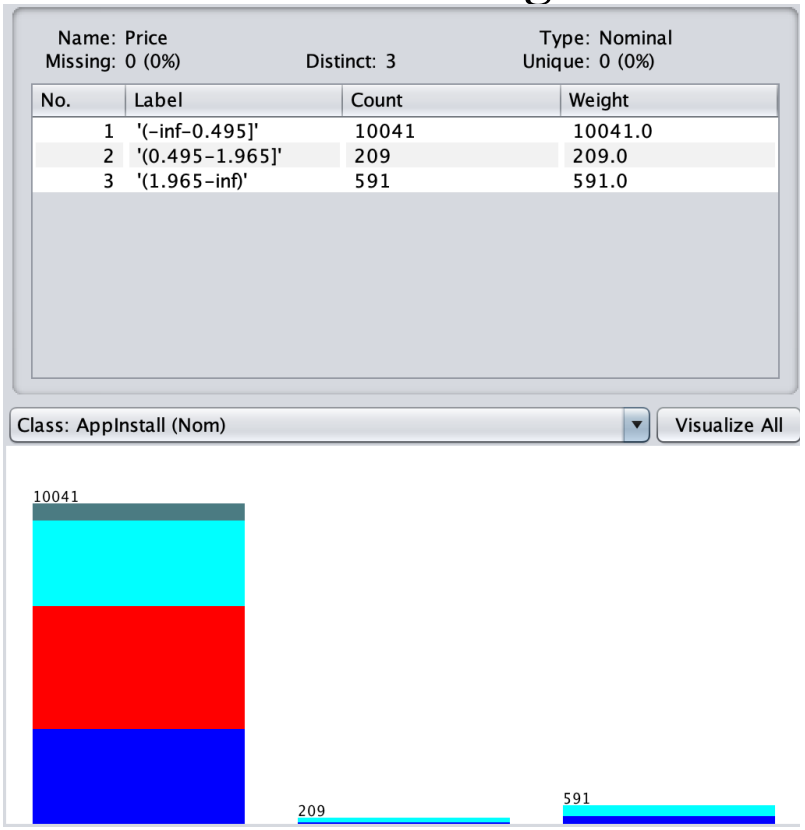
- Random Forest:

```
Correctly Classified Instances        5675               52.3476 %
Incorrectly Classified Instances      5166               47.6524 %
```

Data pre-processing (4/6)

- Bin Size + Installs + **discretize price**
- **Discretize: Weka --> Filters--> Supervised-->Attribute--> Discretize**

Continuous --> Categorical



Data Pre-processing (5/6)

- Bin Size + Installs + **discretize price**
- Naïve Bayes

Correctly Classified Instances	5497	50.7057 %
Incorrectly Classified Instances	5344	49.2943 %

- Random Forest:

Correctly Classified Instances	5678	52.3752 %
Incorrectly Classified Instances	5163	47.6248 %

Data Pre-processing (6/6)

› Bin Size + Installs + discretize price + **resample**

Resample : Weka --> Filters--> Supervised-->Instance--> Resample (with replacement)

Produce a random subsample of a dataset using sampling with replacement.

› Naïve Bayes

```
Correctly Classified Instances        5522               50.9363 %
Incorrectly Classified Instances      5319               49.0637 %
```

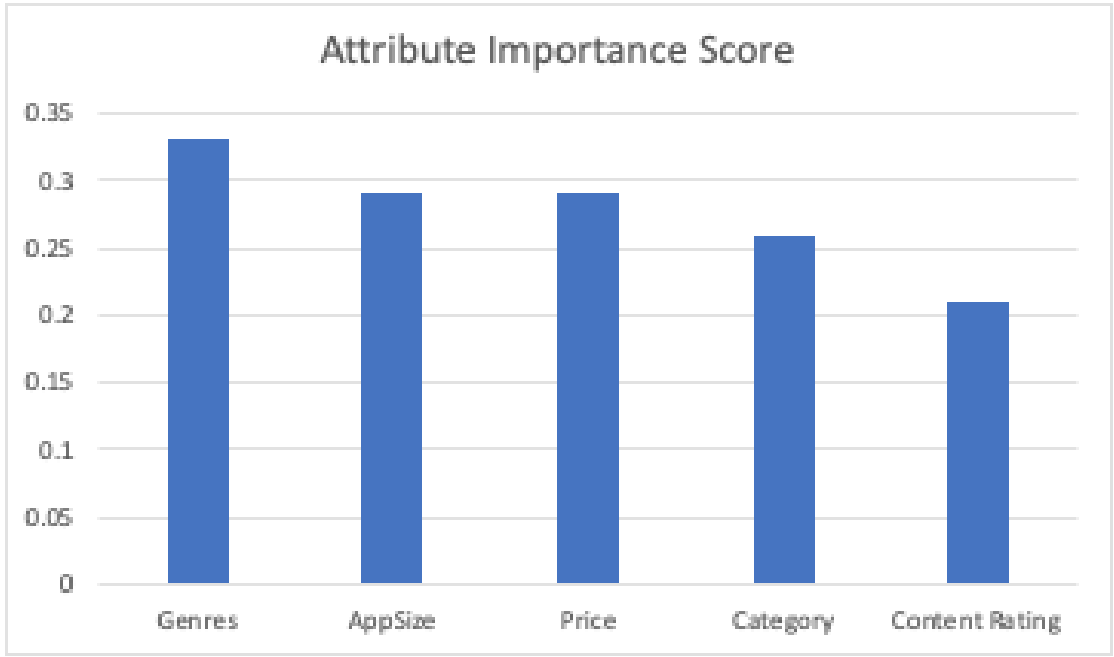
› Random Forest:

```
Correctly Classified Instances        5917               54.5798 %
Incorrectly Classified Instances      4924               45.4202 %
```

Data finding: Attributes Importance(Random Forest)

```
Attribute importance based on average impurity decrease (and number of nodes using that attribute)

      0.33 (   537)  Genres
      0.29 (  7564)  AppSize
      0.29 (  8332)  Price
      0.26 (  1833)  Category
      0.21 (  8559)  Content Rating
```



Data finding

› We prefer underpredict than overpredict the range of installs prediction.

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1210  778 1234   19 |    a = 10,000+
  650 2491  624  107 |    b = 1,000,000+
  725  426 2025   13 |    c = less than 10,000
   24  293   31  191 |    d = 100,000,000+
```

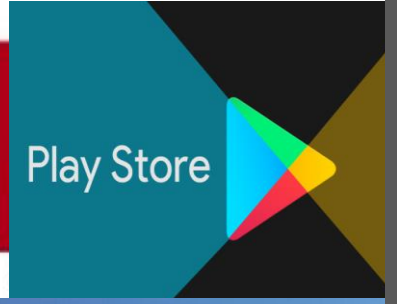
Click to add text

Minimize Error:
C classified as A, B, D
A classified as B, D
B classified as D

Underpredict %
A: 38.07%
B: 32.9%
D: 64.56%

Overpredict %
C: 36.5%
A: 24.59%
B: 2.76%

Conclusion and way ahead

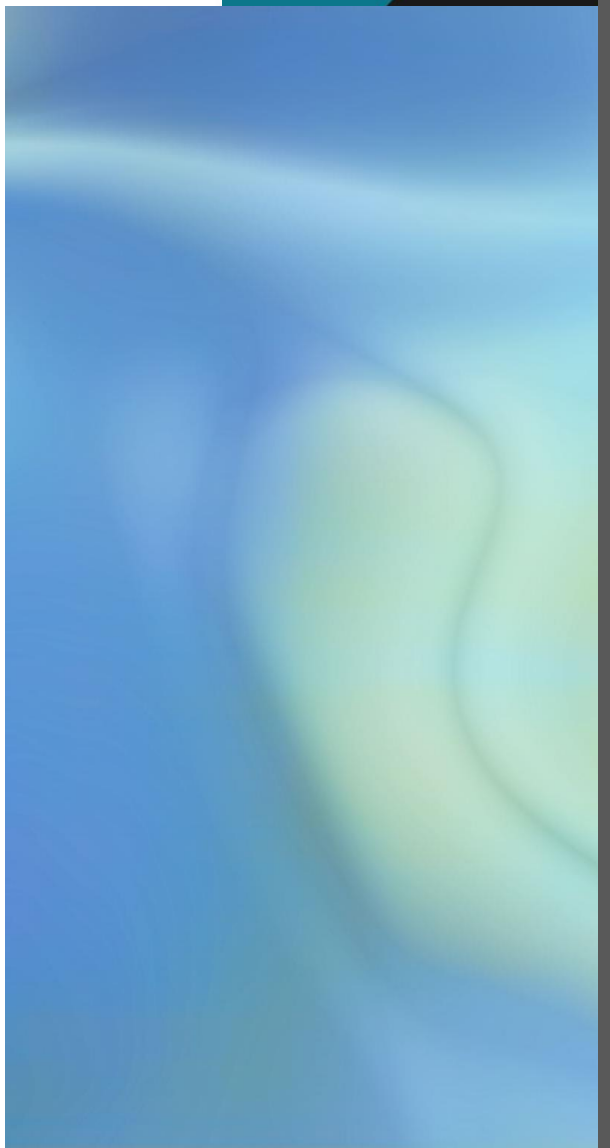


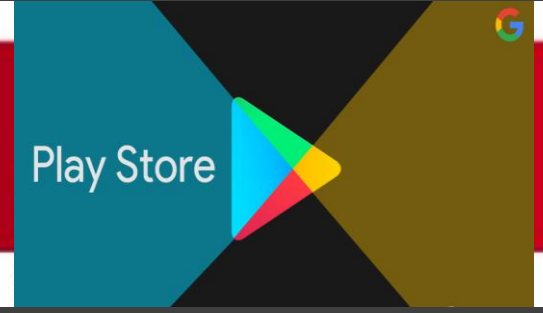
Direct Benefit

- Give developers the prediction range of installs

Indirect Benefits

- Strategy plan for the apps
- Possibility of enhancing application performance





References

https://www.kaggle.com/lava18/google-play-store-apps?select=googleplaystore.csv

http://www.bluecloudsolutions.com/2015/02/06/how-many-downloads-should-my-app-get/

Acknowledgements

Thanks to Professor Mingdi Xin, Ziyi Cao, Jooho Kim, Ming Gu for continuous guidance and support


Thank
You!