MGMT FE 273 Final Project Report: Google Play Store

By: Balaji Chandrasekaran, Edward Kim, Kevin Kung, Aravind Krishnamoorthy, & Kevin Young

EXECUTIVE SUMMARY

Over the past ten years, mobile device apps have become a large industry unto itself. Google's app store, called Google Play Store, has been at the forefront of this industry, with millions of apps available today. But what makes an app successful? It is this question that our team investigated. We used datasets on Google apps (source: Kaggle) from 2010 through 2018. The first dataset had 10,481 records on different mobile apps available on the Google Play Store, with 13 different attributes. The second dataset had 64,296 records of mobile app reviews. The two datasets were scrubbed for incomplete, missing, or skewed data, then merged together.

Using exploratory data analysis, we were able to derive business insights and visualize them from the dataset. We were also able to interpret relationships between different attributes. Next, we pursued a predictive model that would help us determine whether several key attributes could be used to accurately predict if the app is free or paid. The attributes used as the inputs in the model were number of installs, number of reviews, ratings, category, app size, and user sentiment. The final machine learning model yielded the best combination of accuracy (88.3%), precision (62.7%), and recall (57.5%) for solving the revenue generation business problem

Key takeaways and recommendations followed running the model. For specialized (e.g., medical, business, finance) apps, it is critical to have excellent ratings, while price can be at a premium level (up to $80) and business should target providing high quality service to niche users. For non-specialized apps, small file size, free or low price, and above average ratings are important. The key business drive should be to hit more than 1 million users and revenue generation can be targeted through advertisements or in-app purchases. Many other useful

insights can be drawn from the analysis, to the benefit of the app developer community in search of hit app ideas.

Business Idea, why it's important, and the Data Source

One of the biggest and fastest growing industries in the business world, the mobile app ecosystem turned 10 this past summer for both the Apple App Store and the Android Market. This ecosystem encapsulates millions of app developers as well as billions of smartphone owners who use these mobile apps in their everyday lives. The major distribution channel for mobile apps is an app store. An app store (or app marketplace) is a type of digital distribution platform for a smartphone, tablet, and computer software.

Apps provide many benefits. They have the ability to reinforce a brand's visibility and accessibility with their quick customizable user experiences. Functionalities like location tracking, social integration, and push notifications help companies connect with the growing mobile user base. Additionally, having an app can create a direct marketing channel between the company and the consumer by building a database of prospects/clients.

Mobile apps have the ability to generate revenue either via direct sales or through advertisements. In 2015, the total revenue generated across all mobile operating systems was about $70 billion. In 2016, revenue reached $88 billion, and by 2020 the predicted forecast for combined mobile app revenue will reach a staggering $189 billion[1]. To know which apps are in demand and successful, we need to identify the types of apps currently in the app store, specific apps that are selling, purchasing trends of consumers, popular categories, and features customers like or hate the most.

In this project we are trying to gain insights from the Google App store marketplace to uncover trends and correlations between attribute that makes apps a succesul revenue generator. We specifically chose the Google App marketplace due to its open source environment and the number of apps in comparison to Apple's. As of January 2017, the total number of apps in the Google App store is over 2.7 million and is seeing an 82%[2] growth. The Apple store number of apps during the same timeframe is 2.2 million and has higher restrictions for app implementation. This business idea is important because we can get a snapshot of a higher growth app marketplace and develop a model to quickly see how accurate and successful apps will perform given certain variables. Additionally, by identifying these key attributes we allow app developers to become more competitive through our data driven path. App developers with the new insights can make better decisions with variables like market distribution between different categories, the review rating of apps, and which apps tend to be free or paid. The data source used in this project is from Kaggle, a sample dataset of 10,839 Apps and 64,296 review data in the Google Play Store from 2010-2018.

Data Summary, Description, and Visualization

The dataset available from the source were split into two parts. Dataset 1 contains 10,481 records with information regarding the applications available on Google Play. Each record is described by 13 attributes which are listed as follows:

- App - Application name
- Category - Category the app belongs to ss
- Rating - Overall user rating of the app
- Reviews- Number of user reviews for the app
- Size - Size of the app
- Installs - Number of user downloads/installs for the app
- Type - Paid or Free
- Price - Price of the app
- Content Rating - Age group the app is targeted at - Children / Mature 21+ / Adult
- Genres - An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.
- Last Updated - Date when the app was last updated on Play Store
- Current Ver - Current version of the app available on Play Store
- Android Ver - Min required Android version

Dataset 2, with 64,296 records contains the top 100 relevant reviews of each application. The following describes the information contained in this dataset:

- App - Name of app
- Translated_Review - User review
- Sentiment - Positive, NaN, Negative, Neutral
- Sentiment_Polarity - Sentiment polarity score
- Sentiment_Subjectivity - Sentiment subjectivity score - measure the number of adjectives

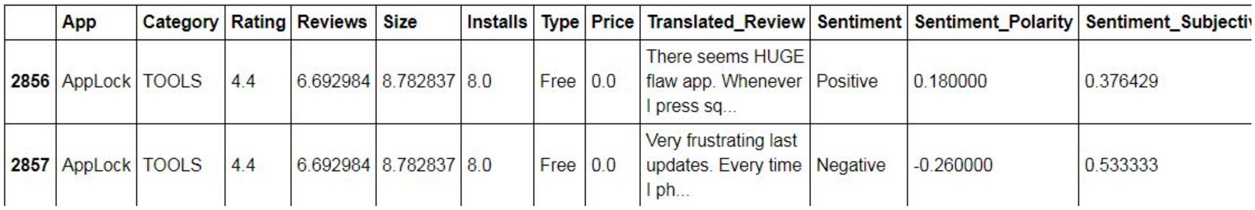
Data Wrangling

Dataset 1

The raw data set so obtained were processed to eliminate duplicates. A total of 1181 duplicate records were eliminated from dataset 1. NaN under the rating attribute were replaced by average of its corresponding categories. Size fields were normalized to “mega bytes” basis. Size column containing 'Varies with device‘ description were replaced by Category Mean. Since number of Installs and Reviews had a very high dynamic range: few hundreds to billions, they were transformed to log10 scale. All junk apps with very high prices (> $100) were discarded and the cleaned up data were converted to .ARFF format for model building.

Dataset 2

Every app contains number of positive/ number of negative and number of neutral reviews, Mean score and Mean subjectivity score. The final cleaned up dataset 2 was then merged with dataset 1 using “App Name” as the primary key from Dataset 1 using “inner” join on Dataset 2. This dataset had only 25% apps coverage and only about 1% of apps on “Paid” Apps category. So we limited our analysis to doing basic NLP techniques and looking at overall positive or negative score percentiles across different categories. (Appendix I). Figure below

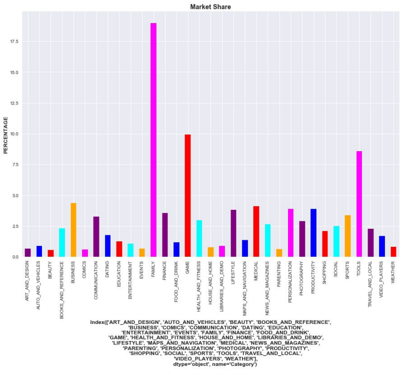


	App	Category	Rating	Reviews	Size	Installs	Type	Price	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjecti
2856	AppLock	TOOLS	4.4	6.692984	8.782837	8.0	Free	0.0	There seems HUGE flaw app. Whenever I press sq...	Positive	0.180000	0.376429
2857	AppLock	TOOLS	4.4	6.692984	8.782837	8.0	Free	0.0	Very frustrating last updates. Every time I ph...	Negative	-0.260000	0.533333

illustrates of a merged dataset 1 and 2.

Data Visualization and Business Insights

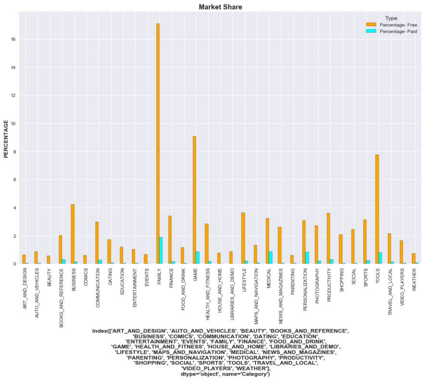
Market Share

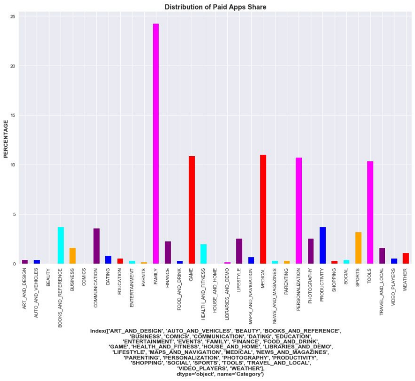


Firstly, the market share of all apps from the dataset is presented. Of the 34 categories, 37.5% of apps belonged to 3 categories namely Family (17.5%), Games (10%), and Tools (8.5%). 28% of market share is more or less evenly distributed among Business, Communication, Finance, Lifestyle, Medical, Personalization, Productivity, and Sports categories.

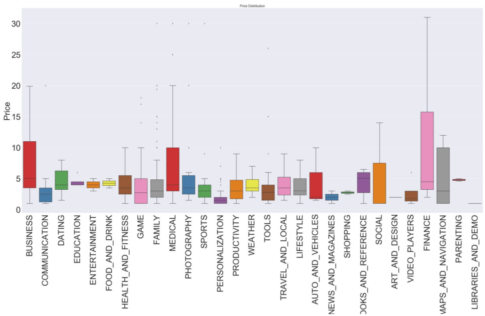
Market Distribution

It can be seen from the market distribution that the paid apps in each of the categories occupy less than 2% of the total share. The largest percentage of paid apps belong to the category “Family”. Of all the paid only apps, Family category takes up nearly 25% of the share, followed by Game, Medical, Tools, and Personalization with each of the category occupying about 10% of the total paid only apps.The mean price of all paid only apps were $22.





Price Distribution

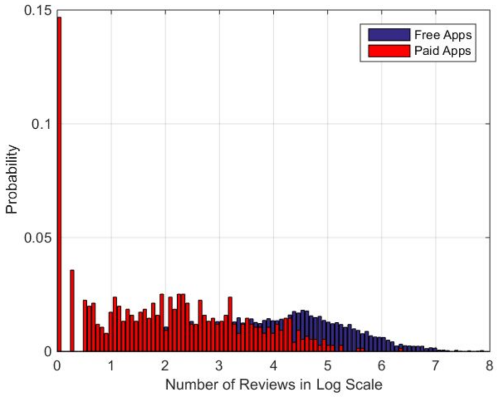


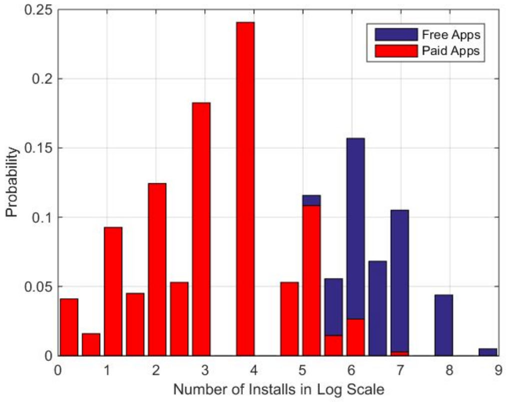
A summary of mean price distribution of the top 10 paid apps are shown here. These categories are so called “specialized’ apps because of their functionality and hence priced higher compared to more “non-specialized common” apps.the specialized apps have mean costs of greater than $5 and are available in categories such as Finance, Business, Medical, Lifestyle, Photography, and Maps.

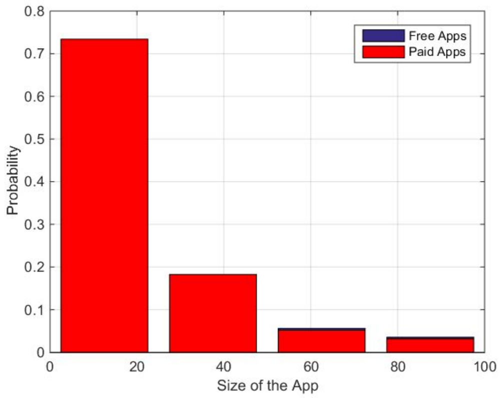
Distribution of Attributes

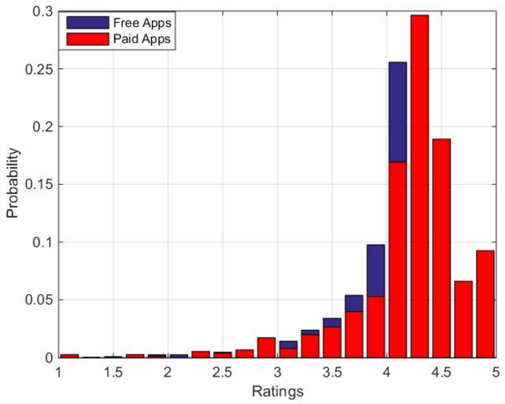
Insightful information regarding some of the variables are given in the following table:

Variable Description	Mean of Variable FREE APPS	Mean of Variable PAID APPS
Number of reviews	234,270	8,800
Number of installs	8,432,439	75,879
Size of apps	20 MB	20 MB
Ratings of Apps	4.165	4.25



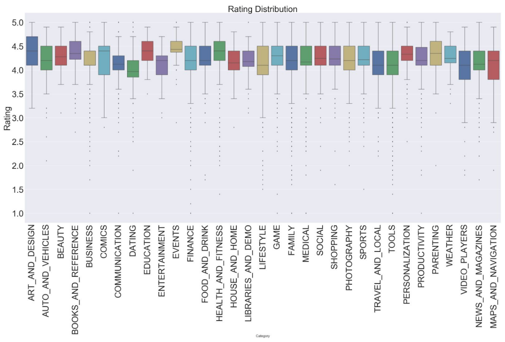






It is seen that the free apps have substantially more installs and reviews. Both free and paid apps have equal mean distribution from the size point of about 20 MB. In comparing to free apps, paid apps have higher mean rating.

Rating Distribution

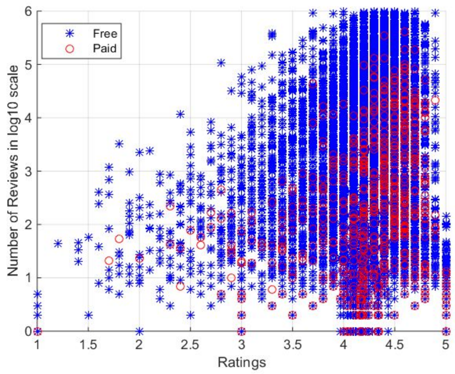


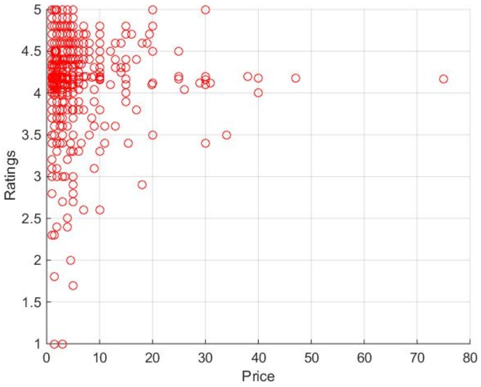
The overall mean rating is found to be 4.125 for all 34 categories including paid and free apps. All Education and Beauty apps are rated above mean and overall has good ratings with lower variance.

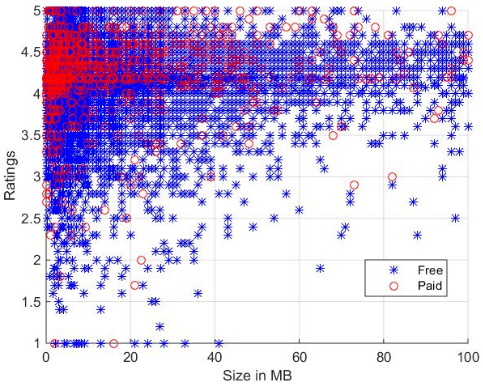
Correlation of Attributes

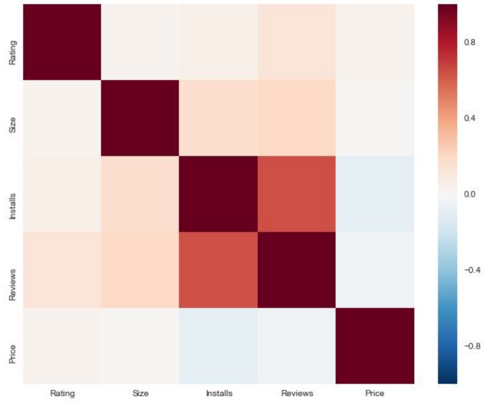
Interesting relationship between variables were observed and are listed/shown below:

1. Highly rated apps have high number of Reviews
2. Highly rated apps are price under $10
3. Highly Rated Apps are within 20 MB (0-20 MB is an ideal App Size)
4. Number of Installs and Number of Reviews received are positively correlated (Rho = 0.62)
5. Number of Installs and Price are negatively correlated (Rho = -0.41)
6. Number of Reviews and Price are negatively correlated









Top Trending Apps Insights

A table of the top trending apps are provided below. All the apps listed below have the following characteristics compared to the means identified in the above frequency charts.

- Have higher Installs
- Higher Reviews
- Higher than 4.125 (Mean Rating)
- Free/Lower Pay
- Lower Size (Gaming apps being an exception)

App Name	Category	Rating	Reviews	Size MB	Installs	Pay Type
Facebook	SOCIAL	4.1	78158306	15.98409	1000000000	Free
WhatsApp Messenger	COMMUNICATION	4.4	69119316	11.30743	1000000000	Free
Instagram	SOCIAL	4.5	66577313	15.98409	1000000000	Free
Messenger – Text and Video Chat for Free	COMMUNICATION	4	56642847	11.30743	1000000000	Free
Clash of Clans	GAME	4.6	44891723	98	100000000	Free
Clean Master- Space Cleaner & Antivirus	TOOLS	4.7	42916526	8.782837	500000000	Free
Subway Surfers	GAME	4.5	27722264	76	1000000000	Free

YouTube	VIDEO_PLAYERS	4.3	25655305	15.792756	1000000000	Free
Security Master - Antivirus, VPN, AppLock, Boo...	TOOLS	4.7	24900999	8.782837	500000000	Free
Clash Royale	GAME	4.6	23133508	97	100000000	Free
Candy Crush Saga	GAME	4.4	22426677	74	500000000	Free
UC Browser - Fast Download Private & Secure	COMMUNICATION	4.5	17712922	40	500000000	Free
Snapchat	SOCIAL	4	17014787	15.98409	500000000	Free
360 Security - Free Antivirus, Booster, Cleaner	TOOLS	4.6	16771865	8.782837	100000000	Free
My Talking Tom	GAME	4.5	14891223	41.866609	500000000	Free
8 Ball Pool	GAME	4.5	14198297	52	100000000	Free
DU Battery Saver - Battery Charger & Battery Life	TOOLS	4.5	13479633	14	100000000	Free
BBM - Free Calls & Messages	COMMUNICATION	4.3	12842860	11.30743	100000000	Free
Cache Cleaner-DU Speed Booster (booster & clea...	TOOLS	4.5	12759663	15	100000000	Free
Twitter	NEWS_AND_MAGAZINES	4.3	11667403	12.470189	500000000	Free

Model Building

The objective of the machine learning model is to build a tool that helps to predict how to price a given app whether to continue it as free and make revenue from the advertisements and in app purchases or to make it as paid app. In this model, features like Category, Ratings, Number of Reviews used, Size of the application and Number of Installs are used for classifying the type of the application. We discarded features like Last updated, Current Version and Android version since they are irrelevant to the problem. The given dataset is imbalanced with about 93% free apps (8902 instances) and only 7% contains paid apps (756 instances)

Feature/Attribute Selection

Features like Category, Ratings, Number of Reviews used, Size of the application and Number of Installs are used for classifying the type of the application. Genres is sub-class of Category and it has been dropped since Category has already been used in the model and it will limit the scope of the analysis to scalable limits. User sentiment and polarity scores are available for only 25% of the apps considered and only 1% of the paid apps category had user sentiment matches. So we discarded this information in the model building. WEKA attribute selection feature is used for gaining insights for picking features.

Top 5 overlapped features from both the chosen methodology are shortlisted and used for model building.

Ranker + Class Attribute Evaluator	**Ranker + Information Gain Attribute Value**
5 Installs 2 Rating 3 Review 4 Size 1 Category	0.06922 1 Category 0.04971 5 Installs 0.02835 2 Rating 0.01536 3 Reviews 0.00911 4 Size

First Iteration

In the first iteration all cleaned up data samples (8902 “Free” and 734 “Paid” are used). This is split into 80% training and 20% testing based on random sample splitting. Performance measurement metric was set for using precision and recall since this is highly imbalanced dataset.

Algorithm	**Accuracy**	**Precision**	**Recall**
Naïve bayesian	87.215	94.74	9.78
Logistic Regression	91.09	25	2.43
Decision Trees (J48)	92.49	75.67	17.07

Confusion Matrix	Label Paid Apps (1)	Label Free Apps (0)
Predicted Paid Apps (1)	30 (TP)	25 (FP)
Predicted Free apps (0)	126 (FN)	1746 (TN)

Though this model has good accuracy, the precision and recall of capturing the paid apps category is very poor in this model. The main flaw is the high imbalance of target variable for training the model.

Second iteration

In this iteration, three approaches was considered for handling the data imbalance problem.

1. Using oversampling filter provided by WEKA
2. Downsampling the Majority category
3. Oversampling by SMOTE technique only on the training dataset

Using Oversampling Filter Provided by WEKA

Resampling filter by WEKA is applied by setting the “biastoUniformClass” to TRUE. This results in an upsampled “PAID” app category from 734 instance to 4839 instance and also some downsampling done to the “Free” app category from 8904 to 4829 instance.

Algorithm	Accuracy	Precision	Recall
Naïve bayesian	70.49	70.89	71.81
Logistic Regression	77.54	77.23	79.12
Decision Trees (J48)	93.79	90.43	96.16

This method provides best outcome in terms of accuracy, precision and recall.But this method of oversampling technique tends to produce samples similar in training and testing. So this filtering is discarded for the final model building.

Downsampling the Majority Category

In this approach all samples from the “Paid” app category are taken and “Free” app category is randomly downsampled by 3 to give about 2968 samples. The total 3702 samples has now about 80% “Free” app category and 20% paid app category minimizing the imbalance from the original distribution. This dataset is further split into 80% training and 20% testing and final performance is evaluated on the testing dataset

Algorithm	Accuracy	Precision	Recall
Naïve bayesian	83.24	65	27.46
Logistic Regression	82.97	62.5	28.17
Decision Trees (J48)	83.37	59.79	40.84

Oversampling Only Training Dataset with SMOTE Filter

In this approach all samples from the “Paid” app category are taken and “Free” app category is randomly downsampled by 2 and merged 1to give about 5236 total samples . This is split into 80% training (4188 samples with 3603 Free App category and 585 Paid App category. Testing dataset contains about 1048 data samples with 899 “Free” app category and 149 “Paid” app category. Scikit learn library “imblearn-over-sampling” was used for generating the oversampled instances. The training dataset was oversamples with k-means nearest neighbor SMOTE filter and the distribution was balanced to about 3600 sample instances of “Free” and “Paid” app category each for training.

Algorithm	Accuracy	Precision	Recall
Naïve bayesian	77.27	68	24.46
Logistic Regression	74.68	65	23
Decision Trees (J48)	86.0	50.5	57.04

Binning Continuous Variable Attributes to Categorical Data:

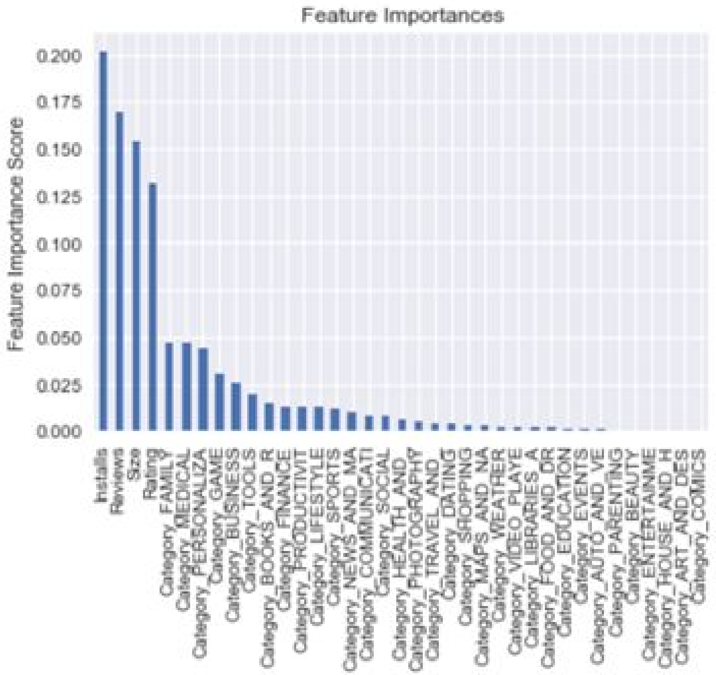
Attribute other than categories are continuous in nature. In this iteration, based on the distribution of data variable width binning strategy was adopted to transform these variables.

Ratings	**Range**
Poor	0 to 2.5
Average	2.5 to 3.5
High	3.5 to 4.5
Excellent	>4.5

Size bins	**Range**
Light Weight	0-20 Mb
Mid Weight	20-40 Mb
Heavy	40-80 Mb
Bulky	>80 Mb

Reviews/Installs bins	**Range**
Low	0 -1000
Average	1000-100000
High	100000 – 1 million
Mature Apps	>1 million

Also sci-kit learn library decision tree model API with same configuration settings as WEKA was used for building the decision tree model. This decision tree parameters are evaluated for their significance or order of importance as given in the below figure:



Based on sorting the importance of decision tree attributes, only Categories like Family, Medical, Personalization and Tools, Games, and Business have significance. The rest of the categories did not contribute any significant information gain to the model. Therefore 34 categories are reduced to just 6 categories to avoid overfitting for noise in the model.

The oversampling technique as described in the previous section was adopted and it was tested on different combinations of continuous and binned categorical variables, and the best performance model in the testing dataset is chosen.

Algorithm	**Algorithm Modifications**	**Accuracy**	**Precision**	**Recall**
Naïve Bayesian	Binned Categories attribute	57.82	20.76	69.80
Decision Trees (J48)	Binned Categories and Size attributes	85.87	50.0	55.70
Naïve Bayesian	Binned Categories and Size attributes	66.79	21.97	52.35

Decision Trees (J48)	Binned Categories Size and Rating attributes	82.26	49.06	42.72
Naïve Bayesian	Binned Categories Size and Rating attributes	50.28	19	76.51
Decision Trees (J48	Binned Categories Rating, Installs and Review attributes	75.68	29	52
Naïve Bayesian	Binned Categories Rating, Installs and Review attributes	55.28	24.67	74.22
Decision Trees (J48)	Binned Categories attribute Install & Size	88.26	62.68	57.53

Final model performance in terms of confusion matrix yielded performance of 88% accuracy, 63% precision and 58% recall.

Confusion Matrix	Label Paid Apps (1)	Label Free Apps (0)
Predicted Paid Apps (1)	84 (TP)	62 (FP)
Predicted Free apps (0)	50 (FN)	758 (TN)

Takeaways & Implications

In the age of "big data," machine learning, and AI, the focus on quality data is often lost in the shuffle. Upon reflection of this project, the theme of data quality is a major takeaway from our experience. Our initial data sets had thousands of instances and many attributes, along with several flaws that had to be addressed. Our data needed to be cleaned of duplicate instances, dropped of unnecessary attributes, resampled, and solved for biases, as well as oversampling. We ran three different iterations before settling on our model. As tedious as this process was, there is no compromise with using quality data. In the end, we were confident that we developed a sufficient model for our business need.

Our business value proposition was to identify the most important attributes of top downloaded apps, and use our model to provide entrepreneurs and businesses a model to properly price an app. By identifying these key attributes we allow app makers to become more competitive in a huge market littered with millions of other apps. Key to building our model was the observation that top rated apps were around 20 MB in size, had an average rating over 4.125 (out of 5), and are lower in price (less than $10). We also found that these observations do not fully apply to "Specialized" apps (i.e. apps specifically for medical, financial, and business needs).

Business Recommendations

Using our model, we have developed recommendations for both specialized and non-specialized apps. For specialized apps, developers must target excellent user ratings. The specialized app needs a minimum rating of 4.5 in order to be a top trending app. Specialized apps can also be priced higher than non-specialized apps. However; the max price is $80. Prices over $80 is generally linked to lower downloads and ratings. This is probably due to consumers finding low value added for the cost of the app. Also, due to the higher price tag, users of specialized apps prefer the app to be ad-free. With non-specialized apps, developers only need to target above-average user ratings (scores of at least 4.125), and either make the app a free download or price it to be less that $10. For both types of apps you want to target a file size of less than 40MB and ideally less than 20 MB. The data did show that for gaming apps, users are more willing to download a game over 40MB in size.

In the competitive landscape of mobile apps, businesses and budding entrepreneurs must be vigilant in keeping up to date with the ever changing trends of consumers. Our analysis provides valuable insights to key attributes, relationships, and trends in the app ecosystem, and our model allows for accurate pricing strategies for specialized and non-specialized apps. It is important to remember, that while our analysis and model can give valuable information today, there is no guarantee that this exact model will work in the future. As the industry continues to grow rapidly, it will no doubt become even more competitive. Our model should continue to be evaluated, invalidated, and updated in order to give businesses a leg up, even if at only the margins.

References

1. https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecast/
2. https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/
3. https://www.datasciencecentral.com/profiles/blogs/dealing-with-imbalanced-datasets
4. https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html

Appendix I

This appendix highlights some basic data analysis done from user sentiments on dataset 2 Since reviews were present only for small percentile of dataset, this report does not infer any strong business insights from these reviews.

Generic inferences from analysis of word clouds of highly rated and poorly rated apps suggests:
Strong Negative

- Paid apps with ads strong negative
- Apps with bugs or frequent crashes and malware

Strong Positive

- Specialized apps that are lighter and user friendly
- Faster and ad free

Category Sentiment	**Negative %**	**Neutral %**	**Positive %**
BOOKS_AND_REFERENCE	8	10	81
BUSINESS	10	20	69
COMMUNICATION	23	18	57
DATING	5	15	79
EDUCATION	9	14	76
ENTERTAINMENT	32	14	52
FAMILY	47	2	50
FINANCE	26	12	60
FOOD_AND_DRINK	19	8	72
GAME	39	3	56
HEALTH_AND_FITNESS	10	10	80
LIBRARIES_AND_DEMO	10	35	53
LIFESTYLE	33	41	25
MAPS_AND_NAVIGATION	26	20	52

MEDICAL	21	26	51
NEWS_AND_MAGAZINES	39	6	54
PHOTOGRAPHY	29	15	54
PRODUCTIVITY	10	28	60
SHOPPING	16	18	64
SOCIAL	34	12	53
SPORTS	24	23	52
TOOLS	17	17	65
TRAVEL_AND_LOCAL	29	20	50
VIDEO_PLAYERS	4	32	64
WEATHER	13	10	76