Team participants:
Aleksander Daniel Veske
Jelizaveta Kuznetsova
Maxim Makarsky

Project repository: https://github.com/morger007/Kaggle-Mobile-Apps

Project D11: KAGGLE-MobileApps

Difference in apps between App Store and Google Play

Identifying your business goals

(For this project my(Maxim Makarskiy) opinion that this step of CRISP-DM is minor and doesn't really matter, howev,er we filled this task with maked up story)

Background

Developers with an app or apps that were originally made for one of the platforms (IOS or Android) and who want to estimate how popular will be a product on the other platform for better understanding what will give more profit promotion and inculcation on one of platforms or partial transition to another platform.

Business goals

1) Analyze market on both platforms to find niches which are most popular on one platform and don't have strong concurrence in other to theoretically create a best selling product increasing income of a company

2) Find difference in trends to estimate how popular will be apps from on platform on other platform

Business success criteria

subjective(difference between profit and loss, popularity or amount of downloads per quarter)

We don't really know what to expect, so there are different criterias:

1) If after our mining, the popularity of the app rises at least +10%.

2) This is subjective, but with provided data from us, we can predict how good the app will be.

3) If profit/downloads +10%, this is success, if profit/downloads -10% don't waste time|effort|money for developing useless apps.

Assessing your situation

Inventory of resources

By now we have two .csv files with data collected from appstore and google play from 2013 to 2015 same data could be collected for future periods if needed

https://www.kaggle.com/lava18/google-play-store-apps

https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps

Requirements, assumptions, and constraints

Data understanding	02.12.2020
Data preparation	06.12.2020
Modeling	10.12.2020
Evaluation	14.12.2020
Presentation	16.12.2020

Risks and contingencies

other homeworks, exams or tests could potentially delay our project

Terminology

CRISP-DM - Cross-Industry Standard Process for Data Mining

Costs and benefits

Zero cost

Potential profit

Defining your data-mining goals

Data-mining goals

1) Determine how certain attributes(rating, size.. etc) affect the popularity of the product(app).

2) Identify patterns in data.

3) Train model, so it can predict on which platform app will be more popular.

4) Train model, so it can predict how popular app will be on both platforms.

Data-mining success criteria

If a trained model is working somewhere near our expectations.

Data understanding

Gathering data

Outline data requirements

We need data that contains following information:

- Name (of app)
- Number of downloads
- Price
- Genre
- User rating
- Number of reviews

Verify data availability

Data is available on kaggle:

- For App Store: https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps?select=AppleStore.csv
- For Google Play: https://www.kaggle.com/lava18/google-play-store-apps

Define selection criteria

Because we searched for our data in kaggle we tried to find most filled datasets for both platforms App Store and Google Play. We didn't aim to find the latest data but we wanted to to find datasets that covered a notable amount of apps.

Describing data

We have four .csv files.

Two files related to Google Play

- googleplaystore.csv (1.3 MB) Each app (row) has values for category, rating, size, number of reviews, installs, type, price, content rating, genres.
- googleplaystore_user_reviews.csv (7.31 MB) has data of user reviews for apps (translated review, sentiment, sentiment polarity, sentiment subjectivity)

Two files related to App Store

- AppleStore.csv (818 KB) (id, name, size, currency, price, number of reviews, number of reviews for current version, user rating, user rating for current version)
- appleStore_description.csv (12.37 MB) (id, app name, size, description)

Exploring data

We will explore our data deeper during the 14th week.

However it's clear that because these two datasets were originally gathered by two different people in different times and didn't actually connect to each other they have some differences.

Verifying data quality

We didn't find data that contains percisece information of what we want. Four csv files above are the most accurate datasets for our goals, however, these files were created in 2013 and now they are outdated so for future use it's better to gather data yourself

Planning your project

Tasks

Data understanding	02.12.2020	-	3 hours each person
Data preparation	06.12.2020	-	6 hours each person
Modeling	10.12.2020	-	9 hours each person
Evaluation	14.12.2020	-	8 hours each person
Presentation	16.12.2020	-	2 hours each person

We did not agree, how much each of us will contribute to each task, MAYBE we will try to make everything together, evenly, where everyone will help the other if needed.
We will create other file in our github with updated information which task who and when will do and how much time were spent for task

Tools

Github, Jupyter Notebook, Python and its libraries (all that we used in our homeworks).