Projeto de
Análise de Dados
Mestrado em Informática
Universidade do Minho
2017

Historical Sales and Active Inventory

https://www.kaggle.com/flenderson/sales-analysis

- Each row in the file represents one product.
- The historical data shows sales for the past 6 months.

Order, Just a sequential counter. Can be ignored.

File_type, historic (historical sales) or active (active inventory).

SKU_number, unique identifier of the product in inventory.

Sold_Flag, 1 = sale, 0 = no sale in past six months. Likely the primary target that should drive the analysis!

Sold_count, greater or equal than Sold_flag. Quantity of items sold.

MarketingType, Two categories of how we market the product. This should probably be ignored, or better yet, each type should be considered independently.

New_Release_Flag, Any product that has had a future release (i.e., Release Number > 0)

Release_Number, an identifier of the release (probably not very informative).

ReleaseYear, year in which the product was released.

PriceReg, **LowUserPrice**, **LowNetPrice**, different types of pricing

StrengthFactor, encoding of strength of the product. Possibly the predicted strength of the individual products to sell. May have relation with other variables!

ItemCount, number of items in warehouse.

Hints:

- Classes are binary and highly imbalanced! Discuss the impact in the classification model (e.g. what happens if the model just gives a 50/50 random prediction? what happens if the model just gives a random prediction based on the prior probability of the class?)

- In case of outliers transform values to log scale, e.g. value X can be transformed to log10(X + 1). +1 because if value is 0, log is not existent.

- Train the model on the historic data and predict for the active products.

2015 Flight Delays and Cancellations

https://www.kaggle.com/usdot/flight-delays

airports table

IATA_CODE, An IATA airport code, also known as an IATA location identifier, IATA station code or simply a location identifier, is a three-letter code designating many airports around the world, defined by the International Air Transport Association (IATA).

AIRPORT, full name of the airport

CITY

STATE

COUNTRY

LATITUDE

LONGITUDE

airlines

IATA_CODE, 2-letter code of an airline or identify to which airline a 2-letter code corresponds

AIRLINE, name of the airline

2015 Flight Delays and Cancellations
https://www.kaggle.com/usdot/flight-delays

flights table

```
Data columns (total 31 columns):
YEAR                 int64
MONTH                int64
DAY                  int64
DAY_OF_WEEK          int64
AIRLINE              object
FLIGHT_NUMBER        int64
TAIL_NUMBER          object
ORIGIN_AIRPORT       object
DESTINATION_AIRPORT  object
SCHEDULED_DEPARTURE  int64
DEPARTURE_TIME       float64
DEPARTURE_DELAY      float64
TAXI_OUT             float64
WHEELS_OFF           float64
SCHEDULED_TIME       float64
ELAPSED_TIME         float64
AIR_TIME             float64
DISTANCE             int64
WHEELS_ON            float64
TAXI_IN              float64
SCHEDULED_ARRIVAL    int64
ARRIVAL_TIME         float64
ARRIVAL_DELAY        float64
DIVERTED             int64
CANCELLED            int64
CANCELLATION_REASON  object
AIR_SYSTEM_DELAY     float64
SECURITY_DELAY       float64
AIRLINE_DELAY        float64
LATE_AIRCRAFT_DELAY  float64
WEATHER_DELAY        float64
```

Hints:

- There seems to be some inconsistent airport codes. Check the kernels in kaggle.
- Check the proportion of flights in 2015 that were cancelled or delayed?
- Which airlines have more delays?
- Any particular time of the year with high incidence of delays?
- Use the *merge* function in R to merge table.

IMDB MOVIE DATABASE

https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

color
director_name
num_critic_for_reviews
duration
director_facebook_likes
actor_3_facebook_likes
actor_2_name
actor_1_facebook_likes
gross
genres
actor_1_name
movie_title
num_voted_users
cast_total_facebook_likes
actor_3_name
facenumber_in_poster
plot_keywords
movie_imdb_link
num_user_for_reviews
language
country
content_rating
budget
title_year
actor_2_facebook_likes
imdb_score
aspect_ratio

Hints:

- The dataset contains very heterogeneous features. Some of them are incomplete data. Filter the null values and standardize the numerical attributes can be a good starting point.

- Create different version of the dataset (e.g. only numerical attributes). Some data analysis techniques will be more suitable for certain data types.

- Would you find movies with similar impact or would you try to predict the imdb score or the gross value of the movie?