Project B2: Spaceship Titanic
Team: Anastasiia Alekseienko, Gutemberg Schiessl
Repository: https://github.com/G-F-Schiessl/Spaceship-Titanic

1. Business understanding

Background:

We are in the year 2912, where we've received a transmission from four lightyears away and things aren't looking good.

The *Spaceship Titanic* was an interstellar passenger liner launched a month ago. With almost 13,000 passengers on board, the vessel has to transport emigrants from the solar system to three newly habitable exoplanets orbiting nearby stars.

While rounding Alpha Centauri en route to its first destination—the torrid 55 Cancri E—the unwary *Spaceship Titanic* collided with a spacetime anomaly hidden within a dust cloud and the catastrophe happened. Though the ship stayed intact, almost half of the passengers were transported to an alternate dimension. Our task is to predict which passengers will be transported to another dimension and which won't.

Business Goals:

To help rescue crews and retrieve the lost passengers, we have to use any methods necessary to predict which passengers were transported by the anomaly using records recovered from the spaceship's damaged computer system. In future, this may also be used to increase the efficiency of other rescue operations and help to handle the consequences of the catastrophes better.

Business Success Criteria:

Our quantitative success criteria will be the amount of correct predictions for the people that will be transported and not. For this, we will use multiple methods to identify the best model and the best criterias to predict the transportation.

Inventory of resources:

Resources: 2 lowkey data scientists, 2 notebooks, Jupyter for coding and Canva for presentation.
All the data comes from a Kaggle challenge:
https://www.kaggle.com/competitions/spaceship-titanic

Requirements, assumptions, and constraints:

The access to the data is already public to anyone. The data is free and able to use by anyone.
The Deadline to deliver the report is 28.11.2022 at noon.

The Deadline to present the project is 15.12.2022 at 2pm.

Risks and contingencies:

The notebooks could be fried and we would need to use university computers.

The internet at the dorms could be off (or too slow). An alternative is the university internet.

Any apocalyptic situation may occur, leaving the campus destroyed and the city in chaos. In this situation, we are gonna run away and leave the project undone (or perform it in real life).

Terminology:

Data-mining terminology:

- K-nearest neighbor - a classification method that classifies a point by calculating the distances between the point and points in the training data set.
- Extreme Gradient Boosting (XGBoost) - an efficient open-source implementation of the gradient boosting algorithm. It refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems.
- Logistics Regressions - an example of supervised machine learning algorithm. It is used to calculate or predict the probability of a binary (yes/no) event occurring.
- Support Vector Machine (SVM) - one of the most popular Supervised Learning algorithms, which is used with the goal to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.
- AdaBoost (ADB) - an ensemble learning method, which was initially created to increase the efficiency of binary classifiers.
- Decision tree (DT) - A type of data mining technique that builds a model for classification of data. The models are built in the form of the tree structure and hence belong to the supervised form of learning.
- Random Forest (RF) - consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Terminology related to datasets:

- HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
- CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for *Port* or S for *Starboard*.
- Destination - The planet the passenger will be debarking to.
- Age - The age of the passenger.
- VIP - Whether the passenger has paid for special VIP service during the voyage.

- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the *Spaceship Titanic*'s many luxury amenities.
- Name - The first and last names of the passenger.
- Transported

Costs and benefits:

Costs: 2 notebook: $ 2000, bunch of neurons
Colab notebook: $ 0
Canva: $ 0
Internet: $ 7
Benefits:
Human Life
If we imagine that this situation may really happen, each correct prediction may save about $250.000-500.000. This is the cost of transportation to space per 1 human.

Data-mining goals:

Models: Extreme Gradient Boosting (XGBoost), Logistics Regressions, K-nearest neighbor (kNN), Support Vector Machine (SVM), AdaBoost (ADB), Decision tree (DT) and Random Forest (RF)
Report: for the report the presentation and file with our predictions can be used
Presentation: will be created (as a pdf file) and performed on 15.12.2022
Files to be processed: 2 files with train and test data, create the sample submission file with our predictions

Data-mining success criteria:

The main criteria will be the model that has the best accuracy and precision on the dataset.

2. Data understanding

Gathering data

Outline data requirements

The data contains nominal, ordinal and discrete data. The data set was collected in the year 2912.

We need two data files: train(⅔ of all information, about 8700 samples, 10 columns) and test (⅓ of all information, about 4300 samples, 9 columns and one, “Transported”

needs to be predicted). We would also have the submission file with 2 columns: PassengerID and Status of transportation (in values True or False).

Verify data availability

All the data is available through the following link:

https://www.kaggle.com/competitions/spaceship-titanic/data

In case if we lose the access to it, we may try to contact someone from the Kaggle to ask if they had this data or look for some other datasets (or imagine it by ourselves, relying on true Titanic dataset or some others)

Define selection criteria

The number of datasets, their sources, size and some specifications were described above. The relevant fields are: HomePlanet, CryoSleep, Cabin, Age, VIP, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck and Transported.

Describing data

Our data is a set of personal records recovered from the ship's damaged computer system. The data is separated in two files: train(⅔ of all information, about 8700 samples, 10 columns) and test (⅓ of all information, about 4300 samples, 9 columns and one, “Transported” needs to be predicted). The columns are:

- HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
- CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for *Port* or S for *Starboard*.
- Destination - The planet the passenger will be debarking to.
- Age - The age of the passenger.
- VIP - Whether the passenger has paid for special VIP service during the voyage.
- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the *Spaceship Titanic*'s many luxury amenities.
- Name - The first and last names of the passenger.
- Transported

This data should be preprocessed before we can use it in the models.

We would also have the sample submission file with 2 columns: PassengerID and Status of transportation (in values True or False).

Exploring data

In this spaceship there are people only from 3 planets: Europa, Earth and Mars. There are 201 NaNs in the dataset.

In CryoSleep, the values are booleans. There are 217 NaNs.

In the Destination there are 3: Trappist-1e, PSO J318 and 55 Cancri e. There are 182 NaNs.

VIP is Boolean. There are 203 NaNs.

Transported is Boolean. There are no NaN.

Age goes from 0 to 68 with a few outliers until 79. 50% of the data is concentrated between 13 and 49 years old. There are 179 NaNs.

RoomService, most people don't ask for this service. 65% do not order not even once. There are 181 NaNs.

FoodCourt was extremely close to RoomService. 64% did not order anything. The FoodCourt has an interesting relation with RoomService. It has a negative correlation between them. There are 183 NaNs

ShopingMall, almost 66% of the people did not use the ShopingMall. There is also a negative relation with FoodCourt. There are 208 NaNs.

Spa, 62% did not go to Spa. There are 183 NaNs.

VRDeck, 64% did not use it. There are 182 NaNs.

Verifying data quality

As we examined, data quality is good enough to perform our project, the size of the sets is sufficient, we have already downloaded all the files that we need so we shouldn't have any problems accessing it (even if we lose it on our computers, we can access it online in Kaggle). In case if we will face some problems with the quality of our models' performance, it will possibly occur due to our data processing, not due to the quality of the datasets themselves. It is also possible that for some columns data is not balanced enough, however, we know how to handle this with code, so there should not be any difficulties.

3. Planning the project

Data exploration: Gutemberg: 15h, Anastasiia: 15h
Data Cleaning: Gutemberg: 7h, Anastasiia: 7h
Create the Model: Gutemberg: 8h, Anastasiia: 8h
Test the model in the trainset: Gutemberg: 2h, Anastasiia: 2h
Test the models in the Testset: Gutemberg: 1h, Anastasiia: 1h
Create the sample submission file: Gutemberg: 1h, Anastasiia: 1h
Create Presentation : Gutemberg: 2h, Anastasiia: 2h
Models: Extreme Gradient Boosting (XGBoost), Logistics Regressions, K-nearest neighbor (kNN), Support Vector Machine (SVM), AdaBoost (ADB), Decision tree (DT) and Random Forest (RF)

We are going to use Colab to write the code and Canva to create the presentation. The format of the code file will be .ipynb, of sample submission file .csv, of presentation .pdf.