# Data Scientist Hackathon

# Filters

- **At the beggining we have 1.544 million rows, after apply the first filter(Only one contract), we obtain 1.5 million rows.**

| | CustomerId | Surname | Geography | Gender | HasCrCard | IsActiveMember | EstimatedSalary | application_date | exit_date | birth_date |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 15745584 | EIRLS | Germany | Female | 0.00 | 1.00 | 0.00 | 2018-12-14 | NaN | 1997-09-18 |
| **1** | 14990118 | MOLOCK | Italy | Male | 1.00 | 0.00 | 121,219.28 | 2019-01-08 | NaN | 1980-08-03 |
| **2** | 14733224 | PAWLUCH | Italy | Female | 1.00 | 1.00 | 159,663.59 | 2012-08-01 | 2013-08-09 | 1977-08-19 |
| **3** | 14648573 | NALLS | Spain | Male | 1.00 | 0.00 | 140,827.98 | 2019-06-19 | NaN | 1979-02-27 |
| **4** | 15365443 | EBERLE | Italy | Male | 1.00 | 0.00 | 35,521.28 | 2014-01-26 | 2015-12-04 | 1972-12-21 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1544995** | 14878861 | LEVENSTEIN | Italy | Female | 0.00 | 0.00 | 99,110.94 | 2019-08-26 | NaN | 1990-05-19 |
| **1544996** | 14520120 | HICKERNELL | Germany | Male | 0.00 | 1.00 | 106,807.46 | 2017-12-10 | NaN | 1984-04-08 |
| **1544997** | 14667679 | MAASSEN | Spain | Male | 1.00 | 0.00 | 83,143.54 | 2014-12-19 | 2018-12-10 | 1963-05-09 |
| **1544998** | 14513378 | KENIMER | France | Male | 1.00 | 1.00 | 153,913.74 | 2012-05-25 | 2014-05-03 | 1976-11-09 |
| **1544999** | 15858637 | MANUAL | Spain | Female | 0.00 | 0.00 | 39,908.14 | 2019-08-03 | NaN | 1987-02-16 |

1500000 rows × 10 columns

# Filters

- **After Dropping Italian clients**

| | CustomerId | Surname | Geography | Gender | HasCrCard | IsActiveMember | EstimatedSalary | application_date | exit_date | birth_date |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 15745584 | EIRLS | Germany | Female | 0.00 | 1.00 | 0.00 | 2018-12-14 | NaN | 1997-09-18 |
| **3** | 14648573 | NALLS | Spain | Male | 1.00 | 0.00 | 140,827.98 | 2019-06-19 | NaN | 1979-02-27 |
| **7** | 14523468 | LASKOSKI | Spain | Female | 1.00 | 0.00 | 158,161.23 | 2017-12-28 | 2018-11-19 | 1972-10-30 |
| **10** | 14915115 | MACURA | Spain | Male | 1.00 | 1.00 | 36,090.09 | 2014-12-28 | 2016-05-11 | 1982-09-26 |
| **11** | 15914714 | TAKAOKA | France | Female | 1.00 | 0.00 | 33,775.00 | 2012-04-30 | 2013-08-09 | 1972-11-04 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1544994** | 15002686 | ARANJO | Germany | Female | 0.00 | 0.00 | 171,390.35 | 2012-08-10 | 2016-01-09 | 1978-01-15 |
| **1544996** | 14520120 | HICKERNELL | Germany | Male | 0.00 | 1.00 | 106,807.46 | 2017-12-10 | NaN | 1984-04-08 |
| **1544997** | 14667679 | MAASSEN | Spain | Male | 1.00 | 0.00 | 83,143.54 | 2014-12-19 | 2018-12-10 | 1963-05-09 |
| **1544998** | 14513378 | KENIMER | France | Male | 1.00 | 1.00 | 153,913.74 | 2012-05-25 | 2014-05-03 | 1976-11-09 |
| **1544999** | 15858637 | MANUAL | Spain | Female | 0.00 | 0.00 | 39,908.14 | 2019-08-03 | NaN | 1987-02-16 |

1170743 rows × 10 columns

# Filters

- **After Dropping clients with 2015< contracts**

| | CustomerId | Surname | Geography | Gender | HasCrCard | IsActiveMember | EstimatedSalary | application_date | exit_date | birth_date | Duration |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 15745584 | EIRLS | Germany | Female | 0.00 | 1.00 | 0.00 | 2018-12-14 | 2019-11-30 | 1997-09-18 | 351 |
| **7** | 14523468 | LASKOSKI | Spain | Female | 1.00 | 0.00 | 158,161.23 | 2017-12-28 | 2018-11-19 | 1972-10-30 | 326 |
| **14** | 15165393 | LABIANCA | Spain | Male | 1.00 | 1.00 | 2,612.65 | 2018-02-22 | 2019-06-11 | 1974-07-11 | 474 |
| **15** | 14611239 | DOKKA | France | Male | 0.00 | 1.00 | 72,210.60 | 2019-02-24 | 2019-11-30 | 1986-04-26 | 279 |
| **20** | 15982728 | GOUDEAU | France | Male | 0.00 | 1.00 | 66,465.09 | 2018-02-02 | 2019-06-01 | 1972-12-18 | 484 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1544979** | 15598810 | ABASTA | France | Female | 0.00 | 1.00 | 1,036.11 | 2019-11-12 | 2019-11-30 | 2001-03-06 | 18 |
| **1544982** | 15923060 | EISENSTEIN | Germany | Female | 1.00 | 1.00 | 3,627.11 | 2018-09-24 | 2019-11-30 | 1977-01-21 | 432 |
| **1544983** | 14506236 | MAINETTI | Germany | Male | 1.00 | 1.00 | 2,850.01 | 2018-04-28 | 2019-11-30 | 1988-05-30 | 581 |
| **1544991** | 15067149 | KUBECK | Germany | Male | 0.00 | 1.00 | 91,273.17 | 2019-02-03 | 2019-11-30 | 1982-07-24 | 300 |
| **1544996** | 14520120 | HICKERNELL | Germany | Male | 0.00 | 1.00 | 106,807.46 | 2017-12-10 | 2019-11-30 | 1984-04-08 | 720 |

278571 rows × 11 columns

- **After apply 75% information filter, we mantain the same rows.**
- **After create  2years> contract variable we keep the same rows**

# Combining databases

- **After applying all filters and add the new variables we get:**

| | CustomerId | Surname | Geography | Gender | HasCrCard | IsActiveMember | EstimatedSalary | application_date | exit_date | birth_date | Notcanceled>2 | Age | Products | Balance | Credit_Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 15745584 | EIRLS | Germany | Female | 0.00 | 1.00 | 0.00 | 2018-12-14 | 2019-11-30 | 1997-09-18 | 0 | 21 | 2 | 103,017.85 | 684 |
| **1** | 14523468 | LASKOSKI | Spain | Female | 1.00 | 0.00 | 158,161.23 | 2017-12-28 | 2018-11-19 | 1972-10-30 | 0 | 45 | 4 | 19,947.23 | 630 |
| **2** | 15165393 | LABIANCA | Spain | Male | 1.00 | 1.00 | 2,612.65 | 2018-02-22 | 2019-06-11 | 1974-07-11 | 0 | 43 | 1 | 119,657.24 | 487 |
| **3** | 14811239 | DOKKA | France | Male | 0.00 | 1.00 | 72,210.60 | 2019-02-24 | 2019-11-30 | 1986-04-26 | 0 | 32 | 3 | 65,015.00 | 436 |
| **4** | 15982728 | GOUDEAU | France | Male | 0.00 | 1.00 | 66,465.09 | 2018-02-02 | 2019-06-01 | 1972-12-18 | 0 | 45 | 1 | 43,353.50 | 595 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **278566** | 15598810 | ABASTA | France | Female | 0.00 | 1.00 | 1,036.11 | 2019-11-12 | 2019-11-30 | 2001-03-06 | 0 | 18 | 1 | 66,818.18 | 674 |
| **278567** | 15923060 | EISENSTEIN | Germany | Female | 1.00 | 1.00 | 3,627.11 | 2018-09-24 | 2019-11-30 | 1977-01-21 | 0 | 41 | 2 | 92,435.85 | 522 |
| **278568** | 14506236 | MAINETTI | Germany | Male | 1.00 | 1.00 | 2,850.01 | 2018-04-28 | 2019-11-30 | 1988-05-30 | 0 | 29 | 4 | 124,422.02 | 589 |
| **278569** | 15067149 | KUBECK | Germany | Male | 0.00 | 1.00 | 91,273.17 | 2019-02-03 | 2019-11-30 | 1982-07-24 | 0 | 36 | 1 | 27,271.40 | 745 |
| **278570** | 14520120 | HICKERNELL | Germany | Male | 0.00 | 1.00 | 106,807.46 | 2017-12-10 | 2019-11-30 | 1984-04-08 | 0 | 33 | 1 | -0.00 | 534 |

278571 rows × 15 columns

# Combining databases

- **We obtained the descriptive Statistics of the new variables**

| | CustomerId | HasCrCard | IsActiveMember | EstimatedSalary | Notcanceled>2 | Age | Products | Balance | Credit_Score |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 278,571.00 | 274,005.00 | 274,005.00 | 274,005.00 | 278,571.00 | 278,571.00 | 278,571.00 | 278,571.00 | 278,571.00 |
| **mean** | 15,261,582.32 | 0.51 | 0.82 | 101,129.37 | 0.02 | 37.89 | 2.47 | 79,879.12 | 650.31 |
| **std** | 433,307.65 | 0.50 | 0.38 | 55,538.57 | 0.15 | 11.29 | 1.12 | 56,879.68 | 96.70 |
| **min** | 14,500,002.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | -0.00 | 174.00 |
| **25%** | 14,884,138.50 | 0.00 | 1.00 | 61,105.73 | 0.00 | 31.00 | 1.00 | 33,752.86 | 585.00 |
| **50%** | 15,270,675.00 | 1.00 | 1.00 | 100,158.00 | 0.00 | 38.00 | 2.00 | 77,189.87 | 650.00 |
| **75%** | 15,642,786.00 | 1.00 | 1.00 | 139,368.00 | 0.00 | 45.00 | 3.00 | 119,244.69 | 716.00 |
| **max** | 15,999,996.00 | 1.00 | 1.00 | 357,882.73 | 1.00 | 92.00 | 4.00 | 374,633.66 | 1,000.00 |

# Implementation of the predictive model

- **Why we choose LogisticRegression?**
- **We will use logisticRegression because we don't want to use too much computing power,**
- **We could try using SVC but it give us the same accuracy result.**
- **LogisticRegression allow us to have the probability of canceling the product before 2 years and keep the product after 2 years/cancel the product AFTER 2 years**

**Insights:**

**The insights that we can get are, the model has a excellent accuracy score = 0.98, but this result could be misleading**

**Why?**

**Because most of the Y values are 0. To be precise 2% of the total values of Y are 1 and the rest are 0.**

**That involves that the predictions that we did could be deceptive.**

**We only have to imagine this case: if we have 100 values and we need to predict them according to our model.**

**If we choose 0 for all values, we will get a 98% of accuracy, due to we only made a mistake with 2 values.**

**Anyway, there could be a different solutions but probably they involve that we need to clean the data in a different way**

**Or maybe we should apply different filters.**

```
print(model_1.predict_proba(X_test))
```

```
[[0.97753651 0.02246349]
 [0.98121401 0.01878599]
 [0.96650278 0.03349722]
 ...
 [0.98604953 0.01395047]
 [0.97017238 0.02982762]
 [0.98119315 0.01880685]]
```

```
print(model_1.score(X_test,y_test))
```

```
0.9766310688324509
```