# ABM team - Technical interview

This task is designed to test your skills in analyzing a data set and statistical modelling. You are required to submit all the code you used ­and present gained insights alongside your problem-solving process. You can use one of the following software/programming languages – *SAS, R, Python.*

The file ‘postpaid\_datamart.csv’ contains data for this test. Dataset includes commercial and usage KPIs for postpaid customers of a Telco operator in MENA from October 2015 to December 2015 and churn flag from January 2016 to March 2016. A customer who churned in January has been inactive in the last 90 days (as of end of month) – i.e. has been active last time in October. Goal of the analysis is to analyze provided data and create a statistical model that will **predict churn probability** for each customer. Churn analysis enables the operator to launch proactive actions to retain the customers that are considered valuable.

Below is the data dictionary for the dataset:

|  |  |  |
| --- | --- | --- |
| Field | Data type | Description |
| CUSTOMER\_ID | Int | Customer unique identifier |
| MONTH | String | Month of the data point |
| FREE\_GPRS\_MB | Float | Monthly data usage (MBs) |
| OG\_INTL\_FREE\_CALL\_MOU | Float | Monthly outgoing international minutes of usage (Mins) |
| OG\_ONN\_FREE\_CALL\_MOU | Float | Monthly outgoing national minutes of usage (Mins) |
| TOTAL\_REVENUE | Float | Monthly total revenue from all services.  Total revenue is a sum of tariff plan revenues (not included in this dataset), roaming revenues, revenues from value added services and pay-as-you-go services (not included in this dataset) |
| TOTAL\_REVENUE\_VAS | Float | Monthly revenue from value added services (games, content etc.) |
| ROAMING\_REV | Float | Monthly revenue from roaming |
| F\_CHURN\_3M | Int | Churn flag indicating customer would be active/inactive for the next 90 days from the month of the data point |

Tasks:

1. **[Data manipulation]** - Examine the data structure and import the dataset. Ensure that all data fields have the correct data type, and that no data has been lost during import step.
2. **[Data manipulation]** - Carry out univariate exploratory data analysis on the dataset and answer the following questions:
3. Are missing values an issue in any of the variables? What method have you chosen for missing values imputation and why?
4. Are outliers an issue in any of the variables? What method have you chosen for outliers detection and why?

* Examine ‘FREE\_GPRS\_MB’ variable. Justify your decision with real life examples.

1. Examine and comment on the distribution of each variable (with graphs).
2. **[Data manipulation]** - Carry out multivariate exploratory data analysis on the dataset and answer the following questions:
3. Examine and comment on the seasonality trends (with graphs/charts).
4. Plot pairwise relationship between input variables and calculate pairwise correlations. Do you think any variable needs to be excluded from the model and why?
5. **[Statistical modelling]** - Create a churn prediction model (use logistic regression) and then answer the following questions:
6. How would you evaluate your model’s performance? Define a quantitative measure
7. What variables are important in the model and why?
8. What are the steps you can take to improve the reliability and fit of the model?
   * Which feature engineering techniques could you apply to try to improve existing model quality?
9. **[Statistical modelling]** – Answer the following questions:
10. Identify other data sources, that might be used to improve the model quality.