## CAPSTONE PROJECT

## Problem Statement

## A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month. The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

## Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

|  |  |
| --- | --- |
| Variable | Definition |
| User\_ID | User ID |
| Product\_ID | Product ID |
| Gender | Sex of User |
| Age | Age in bins |
| Occupation | Occupation (Masked) |
| City\_Category | Category of the City (A,B,C) |
| Stay\_In\_Current\_City\_Years | Number of years stay in current city |
| Marital\_Status | Marital Status |
| Product\_Category\_1 | Product Category (Masked) |
| Product\_Category\_2 | Product may belongs to other category also (Masked) |
| Product\_Category\_3 | Product may belongs to other category also (Masked) |
| Purchase | Purchase Amount (Target Variable) |

## Your model performance will be evaluated on the basis of your prediction of the purchase amount for the test data (test.csv), which contains similar data-points as train except for their purchase amount. Your submission needs to be in the format as shown in "SampleSubmission.csv".

## We at our end, have the actual purchase amount for the test dataset, against which your predictions will be evaluated. Submissions are scored on the root mean squared error (RMSE). RMSE is very common and is a suitable general-purpose error metric. Compared to the Mean Absolute Error, RMSE punishes large errors:

## https://datahack.analyticsvidhya.com/static/images/rmse.png

## Where y hat is the predicted value and y is the original value.

## Please note :

## Public leaderboard is based on 30% of the test dataset, while 70% of the dataset is used for Private Leaderboard.

## The final results would be declared only on Private Leaderboard.

## STEPS TO BE FOLLOWED

## 1. Data Cleaning

## If you’re working with Python, [Pandas](https://pandas.pydata.org/) is a great library to use, and if you’re working with R, you can use the [dplyr](https://dplyr.tidyverse.org/) package. Make sure to showcase the following skills:

## Importing data

## Joining multiple datasets

## Detecting missing values

## Detecting anomalies

## Imputing for missing values

## Data quality assurance

## 2.Exploratory Data Analysis

## This is the process of generating questions, and investigating them with visualizations. EDA allows an analyst to draw conclusions from data to drive business impact. It might include interesting insights based on customer segments, or sales trends based on seasonal effects. Often you can make interesting discoveries that weren’t initial considerations. Some useful Python libraries for exploratory analysis are [Pandas](https://pandas.pydata.org/) and [Matplotlib](https://matplotlib.org/). For R users, the [ggplot2](http://ggplot2.tidyverse.org/) package will be useful. An EDA project should show the following skills:

## Ability to formulate relevant questions for investigation

## Identifying trends

## Identifying covariation between variables

## Communicating results effectively using visualizations (scatterplots, histograms, box and whisker, etc.)

## 3.Interactive Data Visualizations

## Interactive data visualizations include tools such as dashboards. These tools are useful for both data science teams, as well as more business-oriented end users.

## For Python users, the [Bokeh](https://bokeh.pydata.org/en/latest/) and [Plotly](https://plot.ly/python/) libraries are great for creating dashboards. For R users, be sure to check out RStudio’s [Shiny](https://shiny.rstudio.com/) package. Your dashboard project should highlight these important skills:

## Including metrics relevant to your customer’s needs

## Creating useful features

## A logical layout (“F-pattern” for easy scanning)

## Creating an optimum refresh rate

## Generating reports or other automated actions

## 4. Machine Learning A machine learning project is another important piece of your data science portfolio. Now before you run off and start building some deep learning project, take a step back for a minute. Rather than building a complex machine learning model, stick with the basics. Linear regression and logistic regression are great to start with. These models are easier to interpret and communicate to upper level management. I’d also recommend focusing on a project that has a business impact, such as predicting customer churn, fraud detection, or loan default. These are more real-world than predicting flower type.

## If you’re a Python user, use the [Scikit-learn](http://scikit-learn.org/stable/index.html) library. For R users, use the [Caret](https://cran.r-project.org/web/packages/caret/caret.pdf) package. Your machine learning project should convey the following skills:

## Reason why you chose to use a specific machine learning model

## Splitting data into training/test sets (k-fold cross validation) to avoid overfitting

## Selecting the right evaluation metrics (AUC, adj-R^2, confusion matrix, etc.)

## Feature engineering and selection

## Hyperparameter tuning

## 5. Communication

## Communication is an important aspect of data science. Effectively communicating results is what separates the good data scientists from the great ones. It doesn’t matter how fancy your model is, if you can’t explain it to teammates or customers, you won’t get their buy-in. Slides and notebooks are both great communication tools. Use one of your machine learning projects and put it into slide format. You could also use a [Jupyter Notebook](http://jupyter.org/) or [RMarkdown](https://rmarkdown.rstudio.com/) file for a communication project.

## Make sure to understand who your intended audience is. Presenting to executives is very different than presenting to machine learning experts. Make sure to hit on these skills:

## Know your intended audience

## Present relevant visualizations

## Don’t crowd your slides with too much information

## Make sure your presentation flows well

## Tie results to a business impact (reduced cost, increased revenue)

## *Stay positive, continue to build projects, and you’ll be on your way to landing a job in data science.*