You are a data scientist working for an e-commerce company. The company is expanding aggressively, and the leadership team has asked you to come up with a plan to increase revenue by 25% in the next quarter without having to cut down on any existing operations.

**1) Problem Definition**

**Convert the business problem into a data problem**

* Providing discounts to existing products to see if more customers can buy.
* Try and see if we can switch a customer to a subscription service than individual buying.
* More Marketing efforts to see if it could help us increase 25%
* Doing more market research to identify what the customers would want to purchase.
* Target a new range of product within a specific age group to see if that could cause an increase in revenue.

**2) Hypothesis Generation**

**Generate a set of hypotheses based on the problem definition**

* Existing customers are more likely to buy the products than new customers.
* Products selling at a discount are more likely to bought in larger quantity
* Getting feedback from existing customers will result in more increased sales and revenue.
* Increasing Delivery speeds should result in happy customers.
* Customers make purchases mostly at weekends and very minimal on weekdays.
* Customer make purchases when they are happy.
* A large number of household buys typically more products.

**3) Data Collection/Extraction**

**What kind of data do you need based on the above hypotheses? Which variables do you require and how would you collect them?**

* Customer information – Name, Age, demographics, salary, location, household size, customer reviews for each product.
* Sales data – various types of products that have been sold, years in which the products have been sold, warehouse size, value of each product sold, date of the product sold

**4) Data Transformation and Exploration**

**a) What kind of visualization techniques will you use to explore the data?**

* A histogram to compare the total number of products sold in various years.
* Box plots to look at products sold to identify any high cost values items which should be marketed aggressively.
* Linear chart to see if product sold has a linear relationship with Customer’s salary, household size.
* Bar charts to visualize which are some of the items most commonly bought.
* Bar chart to see which geography usually brings in the highest revenue.

**b) Do you need to transform any variables before proceeding with the analysis?**

* All categorical variables such as customer location, household size, customer review stars (Doing label encoding to give more priority to bigger reviews)
* Also impute missing values to median for continuous and mode for categorical variables.
* Change date time to datetime variable including finding months and days of the week and year.
* Depending on the relationship between variables, feature engineering can also be performed to eliminate features that does not have an effect with target variable.

**5) Model Building**

**a) What is the evaluation metric for your problem?**

* Since it is a regression problem, mean squared error should be an important metric.

**b) What kind of models will you build?**

* Considering it is a regression problem, we have to see if we have a linear relationship between the revenue increase vs any other factors and sort by linear coefficient to get the values that have a direct correlation with the revenues.
* Random forest tree should also be a good model as it would help us identify correlation factors
* Neural nets might also be a good options as we will be able to compare it with a second model which works on loss functions.

**c) What if your model validation strategy?**

* K-fold validation should also be used to test and validate the model.

**6) Model Implementation**

a) Which model, based on the ones you have built, is best suited to your business problem? Is there any trade-off between the accuracy and the interpretability?

* I’d like to start with simple linear models and see how my K-fold and MSRE performs, based on that I would decide if that is a good model or we need to go deep into random forest or deep learning stuff.
* Simple linear models are always easier to explain compared to something complicated like a neural network, however if we start to realize that linear models are not giving us a good result, we will have to switch to a neural network.

**b) Any specific steps you’ll follow for monitoring your model’s performance?**

* Periodic evaluation of the model MSRE and other validation factors.
* Periodic checks of any anomalies in the model.