**Approach Doc**

**Data wrangling:**

1. Since data is in form of csv file we have to use pandas **read\_csv** to load the data.
2. After loading it is important to check null values in a column or a row
3. If it is present then following can be done,
4. Filling **NaN** values with zero for past **products\_purchased**, using fillna() method as no product where purchased by particular user
5. Filling **signup\_date** column with same lead generation date column **created\_at**, assumed that if someone buys product they have to signup which is same as lead generation date.
6. If Less missing values, we can drop it as well.
7. Changed object data type column to pandas date time columns.
8. Separate date time columns to their respective day, month, years columns.
9. We have also created time difference (**lead\_time\_diff**) column which is difference in days since lead generation to user signed up.

**EDA**:

1. We have perform different types of graphs to check relation between features.
2. Some corelation graphs using heatmaps were also produced.

**Feature Selection:**

1. Finding out the best feature which will contribute and have good relation with target variable using information gain classifier.
2. We select K-best features using tinkering (18 features were selected).

**Model Selection:**

1. We have created a function to play with different classification models.
2. Models we checked with:
3. RandomForestClassifier
4. DecisionTreeClassifier
5. XGBoost Classifier
6. KNN Classifier
7. ANN Classification method (loss: binary cross entropy, early stopping, metrics: Accuracy)

**Model Hyperparameter Tuning:**

1. Choose following method for hyperparameter tuning
2. RandomizedSearchCV --> Fast way to Hypertune model.
3. 3 fold cross validation were used.
4. Assign hyperparameters in form of dictionary.
5. Fit the model.
6. Check best paramters and best score to select the model.
7. Finally ensemble all models by taking mode.