Model Building Checklist:

Business Objective	Evaluate and justify applicability of selected machine learning model against business objective and model use
Modeling Approach	Developer must create a classical statistical model for bench marking purpose
Development	1. 2 OTV are required 2. Standard data quality check applied 3. Raw data should be processed 4. Feature need to make sense 5. Overlap across samples should be avoided
Model Selection	1. Hyperparameter tuning should be documented 2. Document rationale of objective function and evaluation metrics
Model Diagnostic	1. Model interpretation and implementation • Use Partial dependence plot to interpret each input effect on output ("PDP Plot") • Compare each input of overall relationship of PDP versus a relationship of bivariate plot, in place of contradiction • ICE could be useful to uncover heterogeneous efforts. • LIME can be useful for analyzing outlier observation • SHAP has potential to be used to generate Adverse action reasons 2. Evaluate the final selected inputs contribution to the model's predictive power. • Provide relative feature importance based on gain and relative influence, marginal KS can also be considered 3. Correlation analysis for final selected variables 4. Stability analysis for final selected inputs (CSI Analysis)

Stability Analysis:

PDP: can be summarized into three groups:

1. PDP provides strong relationship
   a. PDP and bivariate has same trend
   b. PDP and bivariate has different trend
2. Performance impact analysis to demonstrate the value of keeping these variables into the model
3. Where possible identify others correlated confounding inputs in the model and use alternative method such that 3rd PDP or ICE to supplement the interpretation (For example ICE plot may

show positive trend for half of the population and negative for another half reveal a reasonable interaction effect while PDP will show the flat net effect.

Elaboration Additional ML performance metrics to be reported:

Common Scenario	Objective Function	Evaluation Metric
Binary Classification	Logloss	AUC and KS
Regression	MSE	RMSE

Feature Engineering:

Transformation	Description
_CRLB	Number of times the variable crossed the lower bound of the confidence interval in last 3 months
_R0309	Ratio of average of 3 months / average of 9 months
_D0306	Difference between average of 3 months and average of 6 months
RMA0306	Ratio of max of 3 months to max of 6 months
RMI0306	Ratio of min of 3 months to min of 6 months
RAM0306	Ratio of max of 3 months to average of 6 months
_CRUB	Number of times the variable crossed the upper bound of the confidence interval in last 3 months
_L1	1 month lag value
_L2	2-month lag value
_Rc3	Variable current value/variable 3-month lag value
_Rc6	Variable current value/variable 6-month lag value
_Rc9	Variable current value/variable 9-month lag value
_Rc12	Variable current value/variable 12-month lag value
_Rs3s6	Sum of values of last 3 months/sum of values of 6 prior months
_Rs3s9	Sum of values of last 3 months/sum of values of 9 prior months
_Rs3s12	Sum of values of last 3 months/sum of values of 12 prior months
_Rs6s12	Sum of values of last 6 months/sum of values of 12 prior months
_Rs6s18	Sum of values of last 6 months/sum of values of 18 prior months
_avg3	Last 3 months average
_avg3_CRLB	Indicator that if last 3 months average crossed the lower bound of confidence interval
_avg3_CRUB	Indicator that if last 3 months average crossed the upper bound of confidence interval
_ci_osh	_avg3 – var.mean_CI – var.std_CI, floor with 0
_ci_ush	_mean_CI – _std_CI – _avg3, floor with 0

_cumi_t<j>	(epsilon + current val)/ (epsilon + min value), using last j month time span. Epsilon = 0.000001 to avoid division with 0
_cumx_t<j>	(epsilon + current val)/ (epsilon + max value), using last j month time span. Epsilon = 0.000001 to avoid division with 0
_curr	Current value, value at time t=0
_cv_t<j>	Coefficient of variation using last j month of time span
_div3	(avg. Last 3 month – avg. Prior 21 months)/std. Prior 21 months
_div6	(avg. Last 6 month – avg. Prior 18 months)/std. Prior 18 months
_m2_t<j>	(epsilon + max value)/ (epsilon + min value); using last j month time span. Epsilon = 0.000001 to avoid division with 0
_max_t<j>	Max value using last j month as time span
_mean_CI	Mean of values between months –23 and –3, this is used to define confidence interval of an account
_min_t<j>	Min value using last <j> months
_mm_t<j>	(epsilon + max value)/ (epsilon + min value); using last j month time span. Epsilon = 0.000001 to avoid division with 0
_mx_t<j>	(epsilon + max)/ (epsilon + sum), using last <j> months' time span
_mxme_t<j>	(epsilon + max)/ (epsilon + mean), using last <j> months' time span
_s2_t<j>	(epsilon + sum – min)/ (epsilon + max – min)
_slop_b1_12m	Slope/coefficient of the account level regression with last 12 months as time span
_slop_b1_24m	Slope/coefficient of the account level regression with last 24 months as time span
_slop_b1_24m_12m	Slope/coefficient of the account level regression with the oldest 12 months as time span
_slop_b1_6m	Slope/coefficient of the account level regression with last 6 months as time span
_std_CI	Std of values between –23 and –3, this is used for defining confidence interval of an account
_std_t<j>	Std of values of last <j> months
_sto_t<j>	Stochastic oscillator; (epsilon + current – min) / (epsilon + max – min)
_sum<j>	Sum using last <j> months as time span
_yt_t<j>	2* curr – max – min; using the last <j> months as time span
_dev	Deviation from the mean
_cluster	Using groups from the cluster

Model design document Approach:

1. Model Scope purpose and use
    a. Product/portfolio/population description
    b. Model purpose and intended users
    c. Model Materiality Tier
2. Limitations and compensating controls
3. Model Data

    a. Data Overview
    b. Data sources and reconciliation
    c. Modelling datasets
    d. Sampling methodology and results
    e. Data assumptions and potential weakness
    f. Data Storage
4. Model Specifications
    a. Model objective
    b. Technical summary
    c. Dependent variable
    d. Segmentation scheme
    e. Variable/Model selection
    f. Final model specification
    g. Diagnostic and statistical test
    h. Potential model weakness
    i. Development code
5. Model testing
    a. Testing plan
    b. Testing summary and conclusion
    c. Testing results
    d. Potential performance weakness
6. Justification of modelling approach
    a. Selected modelling approach
    b. Alternate modelling approach
    c. Benchmark and alternative models
    d. Justification of final model
7. Modelling implementation
    a. Implementation overview
    b. Implementation testing
8. Operating and control environment
9. Ongoing monitoring and governance plan
    a. Ongoing monitoring plan
    b. Annual model review

Model Scope and purpose/use:

1. Product/Portfolio/Population overview
    a. Product description, eligible base, responder definition
2. Macro economy industry trends
    a. Unemployment trend
    b. GDP growth rate
3. Model background and purpose:

    a. The purpose of model is to predict the likelihood of customers taking up the product, with objective to determine the customer profiles using deciles and customer level features

Limitation of modelling framework	The model implicitly assumes that the past relationships continue into the future. Such model mayn't project properly given new scenarios which are fundamentally different from past data	Incremental learning model can improve the accuracy of the model. Xgboost models are simplified representation of real-world relationship between dependent and independent variables. The model cannot capture future idiosyncratic events not reflected in model independent variables, shift in portfolio composition, management actions etc.

4. Data sources and reconciliation
    a. Discuss the modelling input data checks, control process to ensure modelling data is properly sourced and is reliable
5. Modelling data vs production data:
    a. Confirm whether the structure and availability of input data created during model development are the same as data available in production environment for scoring.
6. Modelling datasets:
    a. Discuss the process of modelling dataset creation from raw sources, including a flowchart showing how different datasets are merged together to form development datasets
    b. Data Quality check and cleansing: Provide evidence of consistency and integrity checks how data was tested.

Model Data --> Define Dependent Variable --> Feature engineering --> Variable reduction and selection --> hyperparameter tuning --> final model --> Performance testing

KS	X%
KS Change	NA
Rank - Ordering	No rank order break
Usage Cutoff (Decile)	30%
Capture at intended usage cutoff	30% of population capture 70% of responders
MAPE decile level	0.41
Pct Error (overall)	- 0.91%

PSI	NA

7. Model Usage: Provide detailed description of intended model usage and business processes that employ them and usage plan
8. Outlier Treatment: is not necessary as xgboost is a tree-based model. The reason for the same is as follows.
    a. Tree based methods divide the predictor space that is set of possible values for X1,X2,X3 .. Xn into j distinct and non-overlapping regions
9. Descriptive Statistics: Univariate distribution
10. Data Assumptions and potential weaknesses:
    a. Data Assumptions:
        i. The pattern observed in the Dev are roburst and will sustain in future campaigns
        ii. The behavior of natural responders is representative of the behavior of future campaign responders
    b. Limitations of modelling framework:
        i. The models implicitly assumes that past relationships continue into the future
        ii. Such models cannot project properly given new scenarios, which are fundamentally different from history
        iii. xgboost models are simplified representation of real-world relationship between dependent and independent variables. The model cannot capture future idiosyncratic events not reflected in model independent variables, shift in portfolio composition, management actions etc.
11. Data Storage
12. Model Specification
    a. Model Objective
13. Summary Statistics for dependent variable:
    a. Defines Responders and non responders based on business definition

Performance	Performance definition in dev	Dev(#/%)	ITV(#/%)	OTV(#/%)	OTV2 (#/%)
Good					
Bad					
Performance window	Period during which customer respond to the offer	1st July’20 to 31’st July’20			

14. Initial variable reduction process
    a. Zero importance variable removal
    b. Variable step wise reduction
    c. Correlation analysis
15. Final model parameters and saving all the iterations

16. PDP plot for flat curves
17. Check direction of bivariate and pdp plot
18. Diagnostic and Statistical test
    a. Describe model specific statistical test and diagnotic analysis that are conducted to ascertain the conceptual soundness of the model

Model Segment/Component	Type of test	Test Conclusions
	PDP Plots/Bivariate plots	
	PSI	
	CSI	