Table 1

Model	When to use?	Pros	Cons			
SLR	Relation between X and Y is linear and only 1 X is present	very simple, less time for development	Underfitting			
MLR	Y, multiple X and relation between X and Y is linear	Possibility of a good fit	Prone to colinearity			
Ridge	Y ~ X,X1,X2.. relationship is linear, correlation between predictors, when we do not care about feature importance	Reduces error from model caused by colinearity	1. Does not give feature importance 2. Finding best value for alpha is time consuming, not recommended for datasets with huge size unless necessary.			
Lasso	Y ~ X,X1,X2.. relationship is linear, correlation between predictors, when we do care about feature importance	Reduces error from model caused by colinearity, gives feature importance	1. Takes more time for finding features with lesser importance 2. Finding best value for alpha is time consuming, not recommended for datasets with huge size unless necessary			
Decision tree	Y ~ X1 X2 X3..Xn relationship is non linear, combination	1. Skew of data does not matter 2. If pruned then gives good accuracy. 3. Faster than ensemble techniques.	1. Prone to overfitting 2. Less accuracy compared to ensemble models			
Random forest regression	Y ~ X1 X2 X3..Xn relationship is non linear, combination	1. Skew of data does not matter 2. Feature importance 3. Very high accuracy 4. No problem of correlated trees 5. Multiple tree get created parallel, hence it is faster than other ensemble methods	1. Very slow because creates many trees, internally bootstrap sampling also has to be done, for every tree features have to be selected randomly. 2. If cross validation is used, it will further increase processing time to a very huge extent.			
Adaboost regression	Y ~ X1 X2 X3..Xn relationship is non linear, combination, linear	1. Skew of data does not matter 2. Very high accuracy 3. No correlated trees	1. Very slow because creates many trees 2. If cross validation is used, it will further increase processing time to a very huge extent. 3. Serial approach, which makes it even slower.			
KNN regression	Y ~ X1 X2 X3..Xn relationship is non linear, combination , when very high accuracy is not needed, very high processing time also cannot be afforded.	1. Moderate accuracy 2. Fastest algorithm, because it does not lookup entire dataset to make prediction, looks at only neighbours,	1. Not dependable for high accuracy 2. Features need to be scaled 3. Data is looked up as many times as the number of predictions to be made.			

Table 1-1

	KNN	Decision tree	RF	ADB
Dataset size small	Not to be used	Yes	No/Yes	No/Yes
No of predictions small	Yes	KNN will be better	No	Not
Dataset size large	Yes	Yes	No	Not worth unless required high accuracy
No of predictions large	No	Yes	Yes	No problem

Table 1

Model	When to use?	Pros	Cons	
Logistic Regression	Y,X relation is linear, when good predictors are con	High accuracy	If Prediction probability is close in range, predictions are not reliable	
Naive Bayes	Y,X relation is linear, when good predictors are cat	High accuracy, less processing time	Gives lesser acc with con predictors compared to LR	
Decision tree	Y ~ X1 X2 X3..Xn relationship is non linear, combination	1. Skew of data does not matter 2. If pruned then gives good accuracy. 3. Faster than ensemble techniques.	1. Prone to overfitting 2. Less accuracy compared to ensemble models	
Random forest	Y ~ X1 X2 X3..Xn relationship is non linear, combination	1. Skew of data does not matter 2. Feature importance 3. Very high accuracy 4. No problem of correlated trees 5. Multiple tree get created parallel, hence it is faster than other ensemble methods	1. Very slow because creates many trees, internally bootstrap sampling also has to be done, for every tree features have to be selected randomly. 2. If cross validation is used, it will further increase processing time to a very huge extent.	
Adaboost	Y ~ X1 X2 X3..Xn relationship is non linear, combination, linear	1. Skew of data does not matter 2. Very high accuracy 3. No correlated trees	1. Very slow because creates many trees 2. If cross validation is used, it will further increase processing time to a very huge extent. 3. Serial approach, which makes it even slower.	
KNN	Y ~ X1 X2 X3..Xn relationship is non linear, combination , when very high accuracy is not needed, very high processing time also cannot be afforded.	1. Moderate accuracy 2. Fastest algorithm, because it does not lookup entire dataset to make prediction, looks at only neighbours,	1. Not dependable for high accuracy 2. Features need to be scaled 3. Data is looked up as many times as the number of predictions to be made.	

Table 1
