Data Preparation				
	task		Technical Aspect	Tool and Technique
1	Select Data	Window analysis	vintage analysis; Cumulative default Ratio	cumulative distribution countplot
2	Clean Data	Missing data	Leave missing data	Small percentage of missing values could be tolerated Missing values have special meaning and would be treated as a separate category
			Delete missing data	Listwise (complete) or Pairwise Pros: simple and fast Cons: reduce statistical power, problematic on small datasets
			Single imputation	Mean, mode, median; add missing_flag for adjustment; Pros: simple, fast and uses the complete dataset Cons: reduced variability, ignores relationship between variables; not effective where that data contains a large amount of missing values (typically more than 5% of the data)
			Model-based imputation	Regression Pros: simple Cons: reduced variance KNN imputation Pros: imputes categorical and numeric data Cons: performance issue on large datasets Maximum likelihood estimation Pros: unbiased, used the complete dataset Cons: complex Multiple imputation Pros: accurate, cutting-edge machine learning technique Cons: difficult to code without a special function
		Outliers	Detection correction Decision	Tukey criteria, Q-Dixon test, Chauvenet criteria, Grubbs test, Tietjen-Moore test, Extrem Studentized Deviation, Tau test outside ±3 standard deviations, or ±1.5IQR, or 5th–95th percentile range would be labelled as an outlier binning, weights assignment, conversion to missing values, logarithm transformations to eliminate influence of extreme values or Winsorization
		Stationarity	Detection correction	ADF test, KPSS variation
		Multicollinearity	Detection correction	Correlation matrics ; VIF Delete / combination
3	Construct Data	Derived Attributes	univariable	Mathematics function
		Generated Records	multivariables	PCA VAR
4	Integrate Data	Merged Data		SQL-Python/R
5	Format Data	removing illegal characters		
6	Transform data	Discretization	quantitatives variables	WoE; OHE : One-Hot encoding BE : Binary Encoding LE : Label Encoding
		Binning	band-variables	smbinning, optimalbining
		Expert		

7	Select feature	Supervised	Wrapper	Forward selection, Backward elimination, Bi-directional elimination, Exhaustive selection, Recursive elimination
			Filter	Information Gain, Chi-square test, Fisher's Score, Correlation Coefficient, Variance Threshold, Mean Absolute Difference, Dispersion Ratio, Mutual Dependence, Relief, Pearson's correlation, Spearman's rank, ANOVA, Kendall's rank
			Embedded	Regularization, Tree-based methods
		WoE	WoE	Information Value