**Viva – Topics:**

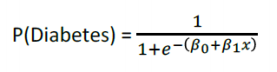
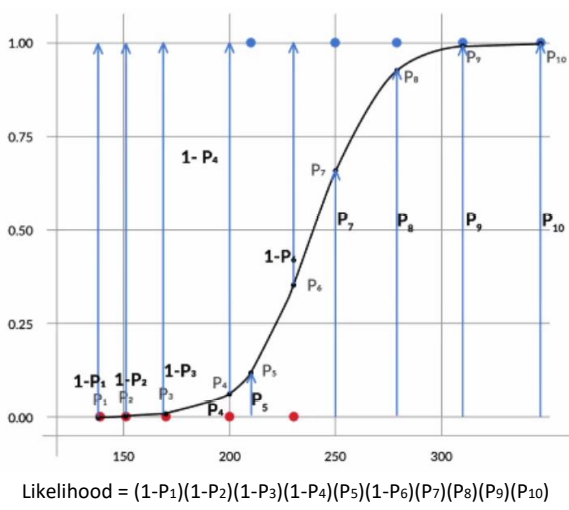
* Data Dictionary
* Missing Value Analysis – how to deal with missing values
* Remove columns with variance
* Outlier Analysis
* Understanding the distribution of the data
* Dummy one-hot encoding
* Binary mapping
* Heatmap
* Packages used for different tasks
* train test split,
* Standard Scaler,
* Logistic Regression,
* RFE,
* stats model
* Feature scaling
* RFE – Recursive Feature Scaling
* What add constant in statsmodels, GLM
* Confusion matrix
* Variance inflation factor
* Model iteration
* P-value
* Formula of accuracy
* How to fit a model
* Sensitivity
* Specificity
* False positive rate
* Positive Predictive Value
* Negative Predictive Value
* ROC Curve
* Finding Optimal cut-off point
* Confusion matrix
* Probability, Sensitivity, specificity
* Precision-Recall trade-off

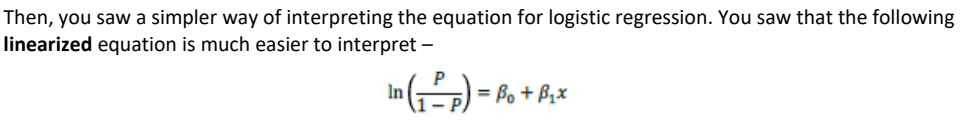
Logistic Regression:

Logistic Regression Notes:

Logistic regression is a supervised classification model. It allows us to make predictions from labelled data, if the target variable is categorical. Initially, binary classification was analyzed. Examples of Binary classification:

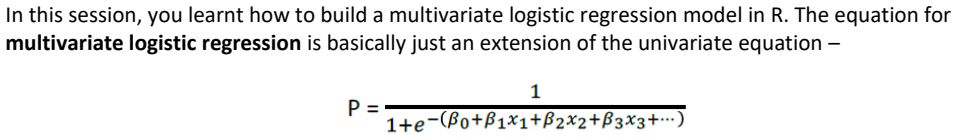
* Bank wants to predict if customer will churn or not
* Machine will break down or not next month
* Whether incoming mail is ham or spam
* Whether person has diabetes or not, based on person’s blood sugar level

A curve that could model diabetes well was the **sigmoid** curve.   
Likelihood: To find the best sigmoid curve, we vary B0 and B1 values to find best fit



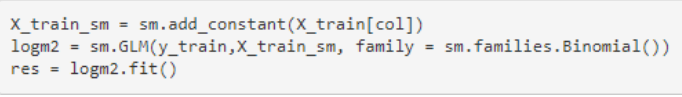
The odds of having probability is (P/(1-P)) and the log odds is ln(P/(1-P)).

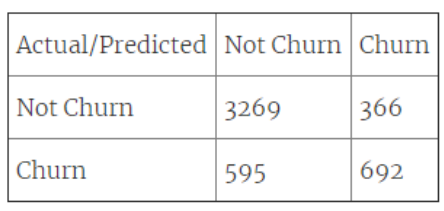
How does odds vary with variation in X?  
With every linear increase in x, the increase in odds is multiplicative.

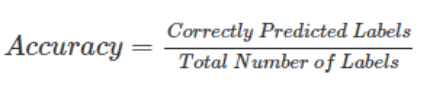
**Multivariate Logistic Regression:**  


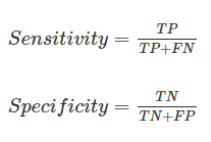
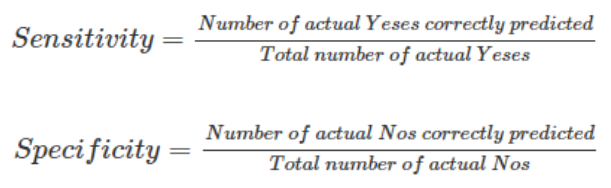
Data Preparation Steps: Missing Value Imputation, Outlier Analysis, Dummy Variable, Test-Split Data, Standardization on scales of data.

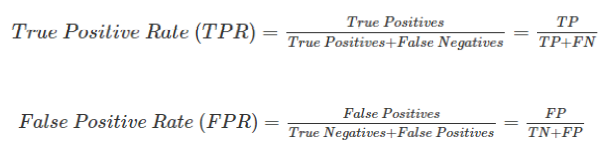
A logistic regression model was built in Python using the **GLM()** under the **statsmodel** library. RFE and then, manual approach using Recursive Feature Elimination (RFE) and then, a manual approach using p-values and Variable inflation Factor (VIF)

  
**Confusion Matrix: Accuracy, Sensitivity, Specificity:**

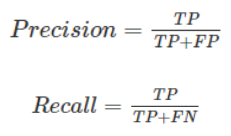








**Model Evaluation: Precision and Recall:**



Sample selection for Logistic Regression (errors):  
1. Cyclical or s**easonal fluctuations** in the business need to be taken care of  
2. The sample should be **representative of the population** on which the model is to be applied in the future  
3. For **rare event sample**, they should be balanced before they are used for modeling

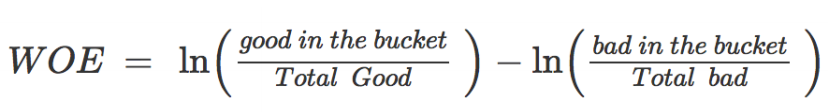
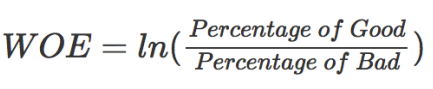
Approaches for transforming independent variables:

* Dummy Variable Transformation
* Weight of Evidence (WoE) transformation
* Continuous variables transformation
* Interaction Variables
* Splines
* Mathematical Transformation
* Principal Component Transformation

**Creating Dummy Variables:**  
For categorical variables of n levels, create dummy variables of n-1 levels.   
For continuous variables, when dummy variables are created via binning, it makes the model more stable  
Disadvantage: If more features are converted to dummy variables, it may result in data clumping

WOE Transformation:

* Calculate woe values for fine binning and coarse binning
* Importance of woe for fine binning and coarse binning
* Usage of woe transformation

Two main advantages of WOE:  
WOE reflects group identity – it captures general trend of distribution  
WOE helps to treat missing values logically for both types of variables – categorical and continuous.  

Once the woe values have been calculated, they should follow an increasing or decreasing trend across bins. If the trend is not monotonic, the bins need to be compressed and woe values need to be recalculated.

**Information Value**:





Information value is an important indicator or predictive value. Mainly, it helps understand how the binning should be done. The binning should be done such that the WOE trend across bins is monotonic – either increasing all the time or decreasing all throughout. But, one more thing that should be taken care of is that IV value should be high.

Model validation:  
Model can be validated on”

* In-sample validation: validated within the same time frame or similar setting
* Out-time validation: validated within a different setting or time frame
* K-fold cross validation: k-iterations on the same dataset to ensure that there is no bias

Model Stability:  
A model is considered stable if:

* Performance stability: Results of in-sample validation approximately match those of out-of-time validation
* Variable Stability: Sample used for model building hasn’t changed too much and has the same general characteristics