Final Project: INDU6611

Author: Shravan Tawri

Project contents

- About the case study and data description
- Handling missing data
- Data preprocessing
  - Handling date/time data
  - Converting categorical data to numerical data
- Method 1: Retention prediction using different classification models
- Feature selection
- Method 2: Classification, but using only most important features
- Method 3: Classification, using data in part B of the case

Case Study

- Scholastic Travel Company (STC) is one of the premium providers of cultural and educational trips.
- David Powell, a data analyst at STC was eager to start a new project. He stressed his new supervisor Stephen Blackford for a new data initiative centering customer retention.
- With retention prediction, he wanted to implement a marketing strategy that would target certain subsets of the client population to save cost and improved yield.

Data description - part A

- Data of 2389 clients from 2009 to 2011
- Data given consists of 56 columns, with each column gives significant values related to client trips.
- Each column in the data sets consists of null values which were found using *‘.isnull()’* function.

Handling NULL (NaN) values

Drop features having NA Values

Loss of data

Used only when majority of data is missing

Replace/Fill NA values

Mean

Median

Mode

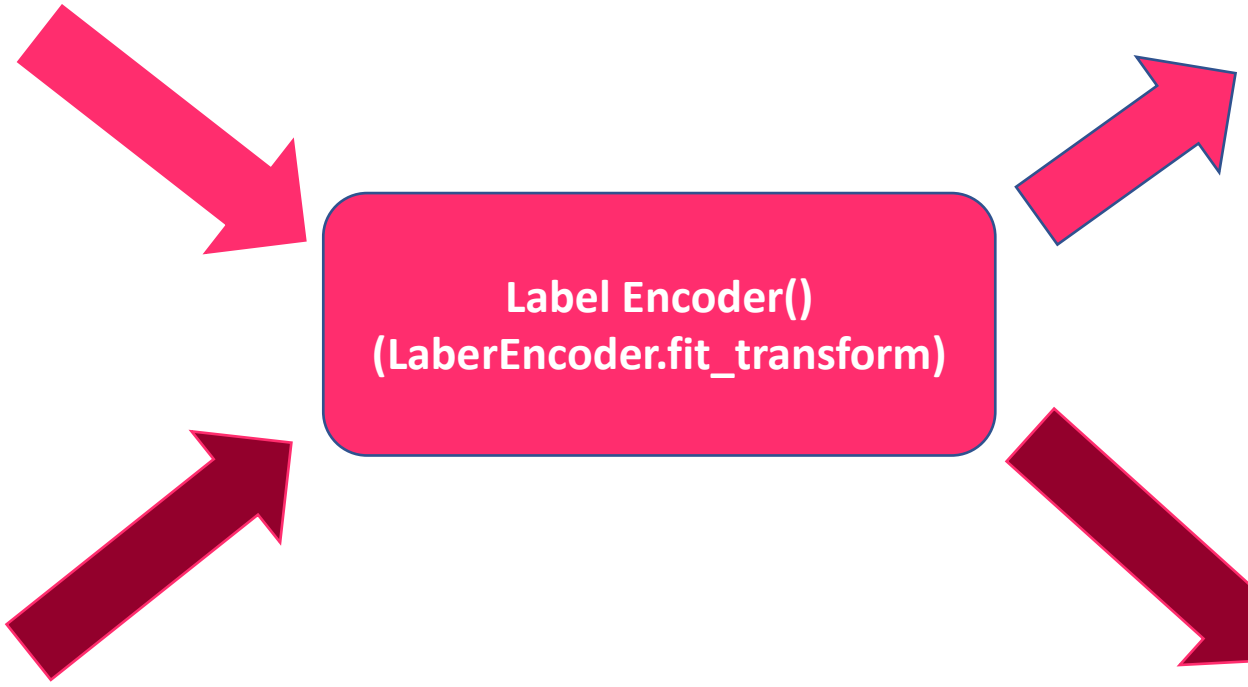
Filling NULL (NaN) values

Continuous data (e.g., FPP: School Enrollment ratio)	Replace with mean
Discrete data (ordered) (e.g., Income level)	Replace with median
Discrete data (non-ordered) (e.g., Travel type)	Replace with mode
Date/Time data	Convert to number of days with departure date being reference and replace NaN with mean

Handling categorical data

Departure Month
January
February
March
April
May
June

Travel Type
Air Travel
Bus
Train

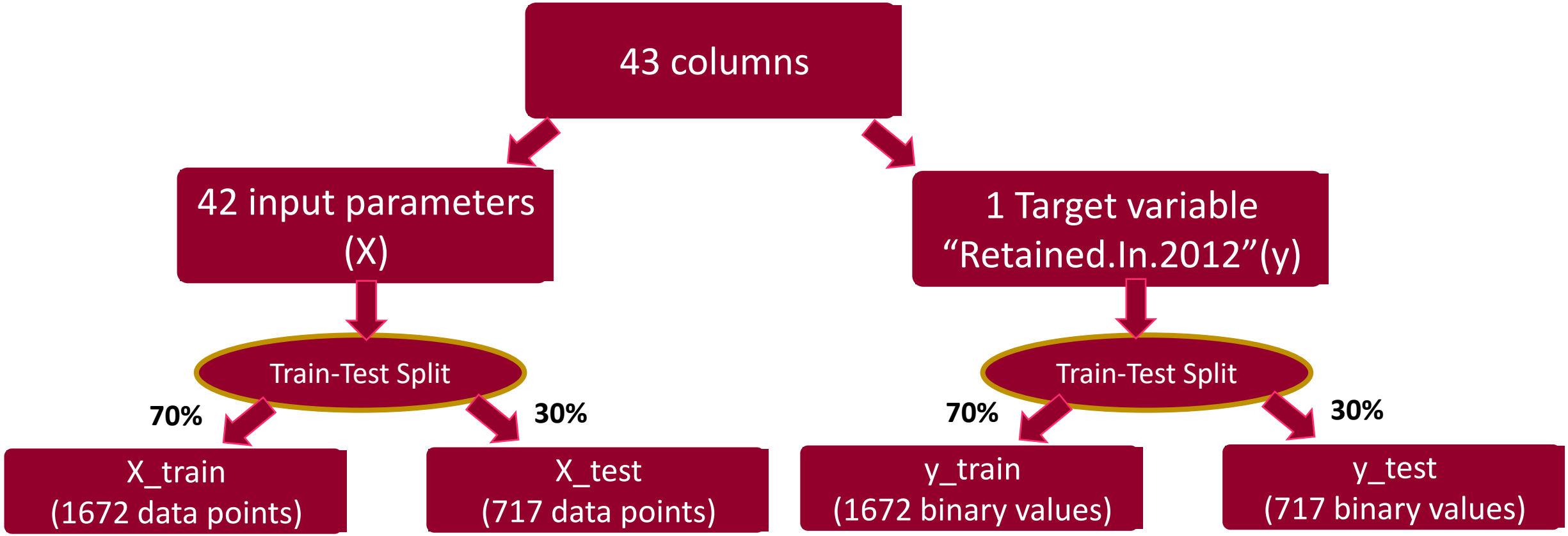


Departure Month
1
2
3
4
5
6

Travel Type
1
2
3

METHOD 1: MODEL BUILDING

- After data cleaning and processing we have 2389 non-null data entries with 43 features (excluding ID number)



Classification models used

- Decision Tree Classifier
- Logistic regression
- KNN Classifier
- Random Forest Classifier
- SVM Classifier

Performance Parameters

- Accuracy
- Confusion matrix

Table of results – method 1

Classifier	Accuracy (%)
Decision Tree	71.68
Logistic Regression	79.21
K-Nearest Neighbor Classifier (k-NN)	61.08
Random Forest Classifier	**80.33**
SVM	60.39

Feature selection using Correlation Matrix

- Used *'.corr()'* function to generate correlation matrix for the features available to check their impact on target variable and their dependence on each other.
- Higher the correlation matrix absolute value, higher is the dependance on target variable.
- Top 5 features are selected (based on correlation values obtained from the matrix generated):

(i) SingleGradeTripFlag, (ii) FPP , (iii) FRP.Active,Total.Discount.Pax , (iv) Is.Non.Annual. , (v) SPR New.Existing

- From the correlation matrix we see that some features are highly correlated e.g FPP and Total PAX, Total discount PAX and No. of non FPP PAX. Therefore, we eliminate one of the similar columns.

METHOD 2: MODEL BUILDING

- Similar to method-1, but we use only the top 5 features for prediction

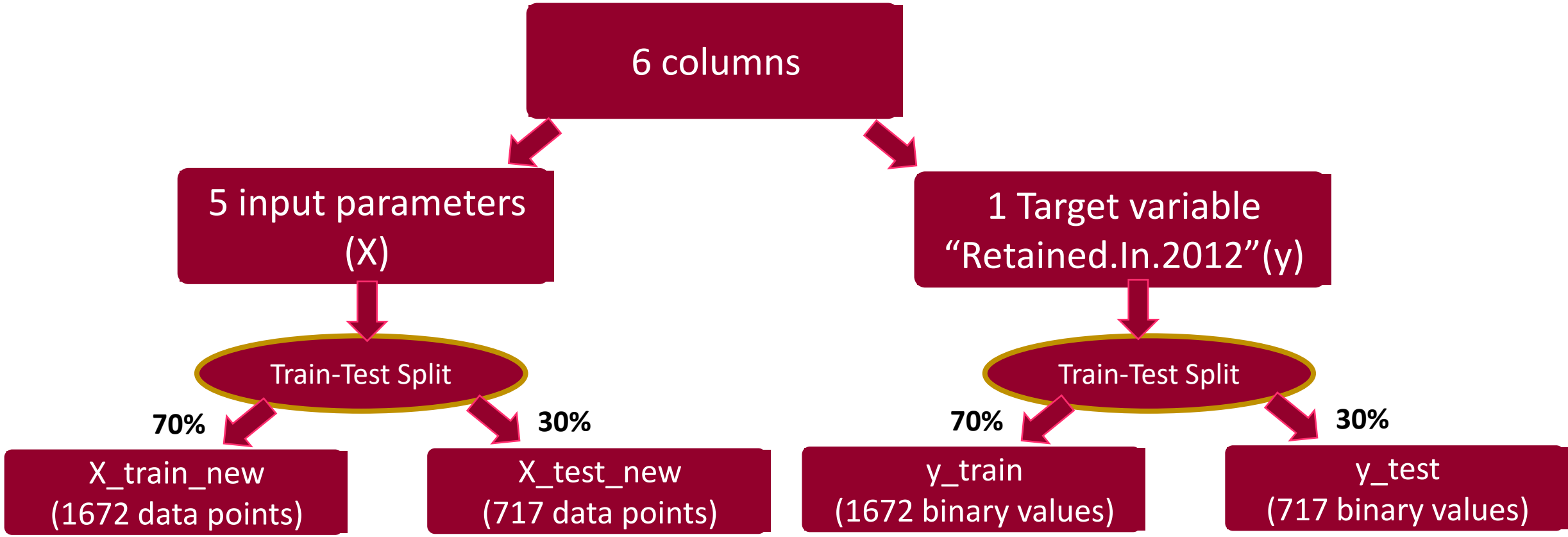
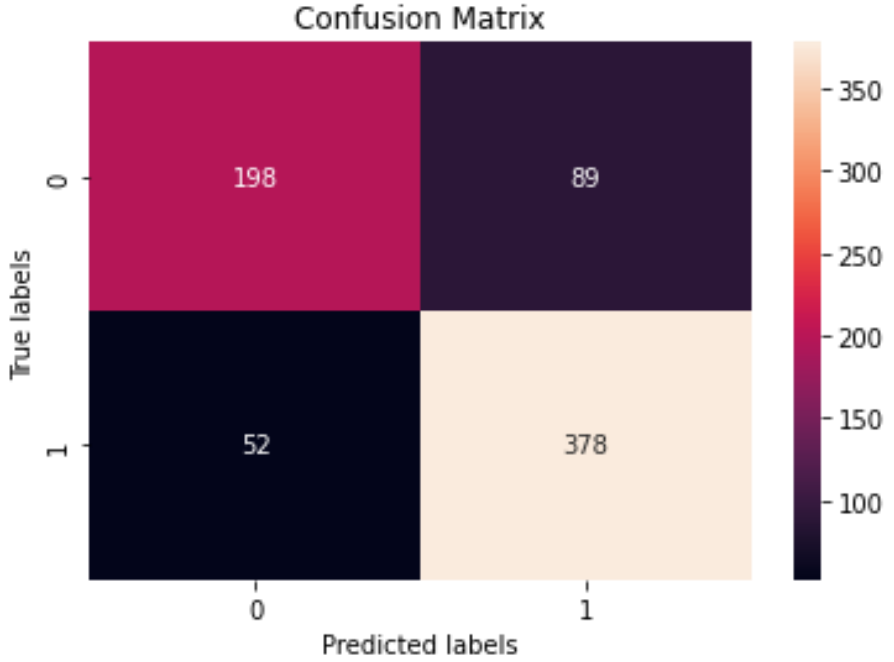


Table of results – method 2

Classifier	Accuracy (%)	Change
Decision Tree	75.59	+5.3↑
Logistic Regression	**80.33**	+13.53↑
K-Nearest Neighbor Classifier (k-NN)	72.52	+11.44↑
Random Forest Classifier	76.42	-3.07↓
SVM	75.31	+14.92↑

Best Model

- The Logistic Regression model trained on data using only the Top 5 features gives the best result with 80.33%
- Since the number of 0s and 1s are fairly equally distributed (40%-60%), the model accuracy and confusion matrix can give a precise representation of the usefulness of our model.



Part B

- Use the Net Promoter Score (NPS) to check if we obtain an improvement in the accuracy.
- NPS ranges from 0 to 10 and is an indicator of a student's level of satisfaction from the trip.

Feature name	NULL (NaN) values
NPS 2011	577
NPS 2010	1229
NPS 2009	1225
NPS 2008	1412
>= 3 FPP Date	9
>= 10 FPP Date	409
>= 20 FPP Date	1024
>= 35 FPP Date	1618

Too many missing values

PART B-1

- Since there are a lot of missing values in most of the columns, only *'NPS 2011'* and *'>=3 FPP Date'* are considered.
- For Part B-1, rows with missing values of the *'NPS 2011'* are dropped. (Disadvantage: loss of data)
- A new feature is added named *'Difference >3 FPP Date'* which indicates the difference in days between the departure date and the date at which the total group size exceeded 3. Missing values are replaced with mean number of days.
- These two new features are then combined with the original data (from Part A) and is used to train classification models.

Part B-1: Model building

- Similar to method-1, but the number of datapoints have now decreased **(LOSS OF DATA!)**

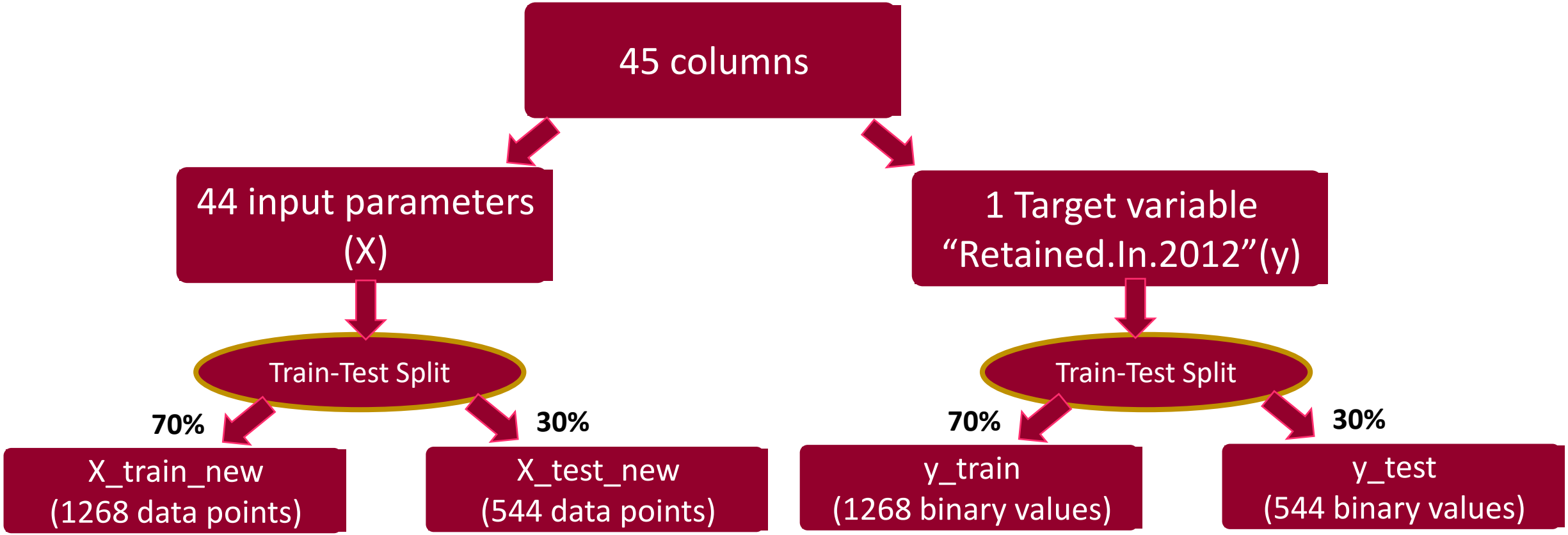
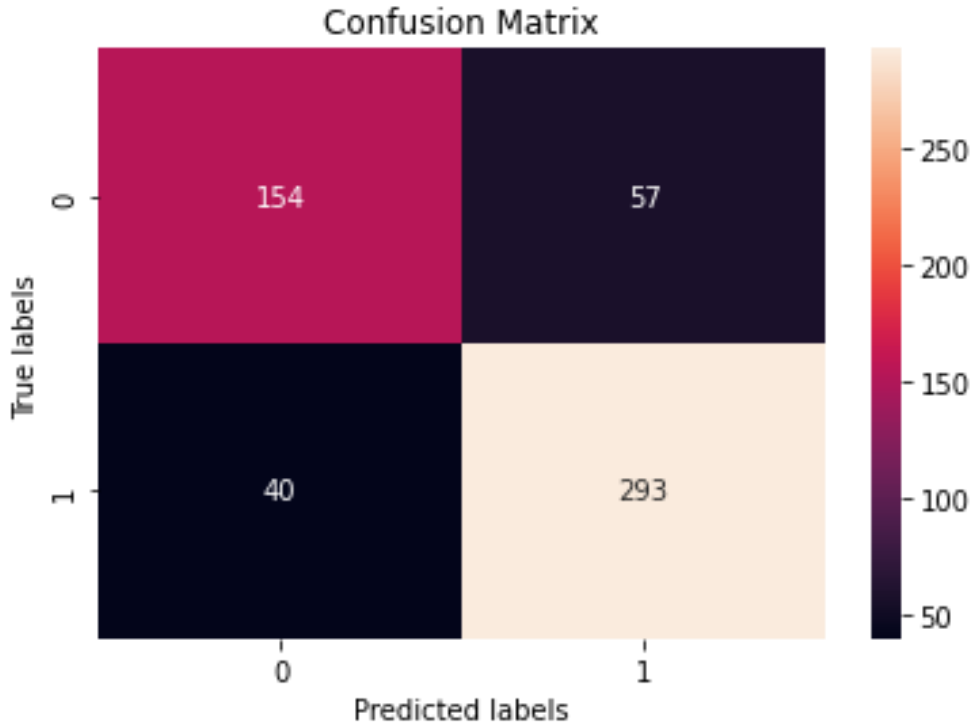


Table of results – Part B-1

Classifier	Accuracy (%)	% Change
Decision Tree	75.91	+0.32↑
Logistic Regression	**82.16**	+1.83↑
K-Nearest Neighbor Classifier (k-NN)	63.41	-9.11↓
Random Forest Classifier	81.06	+4.64↑
SVM	61.39	-13.92↓

Best Model

- The Logistic Regression model trained on data using only the Top 5 features gives the best result with 82.16%
- **PRO** : Improvement in accuracy by ≈2% than Part A.
- **CONS**: We can predict the retention status for students that have a non-NULL NPS 2011 score.



Part B-2 results

- Similar to Part B-1, but instead of dropping rows with NA values, we try filling them with median values of *'NPS 2011'* .
- The accuracy of the model decreases.

Classifier	Accuracy (%)	% Change
Decision Tree	70.85	-5.06↓
Logistic Regression	**79.49**	-2.67↓
K-Nearest Neighbor Classifier (k-NN)	61.08	-2.33↓
Random Forest Classifier	80.33	-0.87↓
SVM	60.52	-13.92↓

Recommendations