E-Commerce Customer Churn Analysis and Prediction

BY: PRIYANKA BAGCHI

PRIYANKA.BAGCHI.PB@GMAIL.COM

SUPERVISOR: DR. SEDEF AKINLI KOCAK, PHD

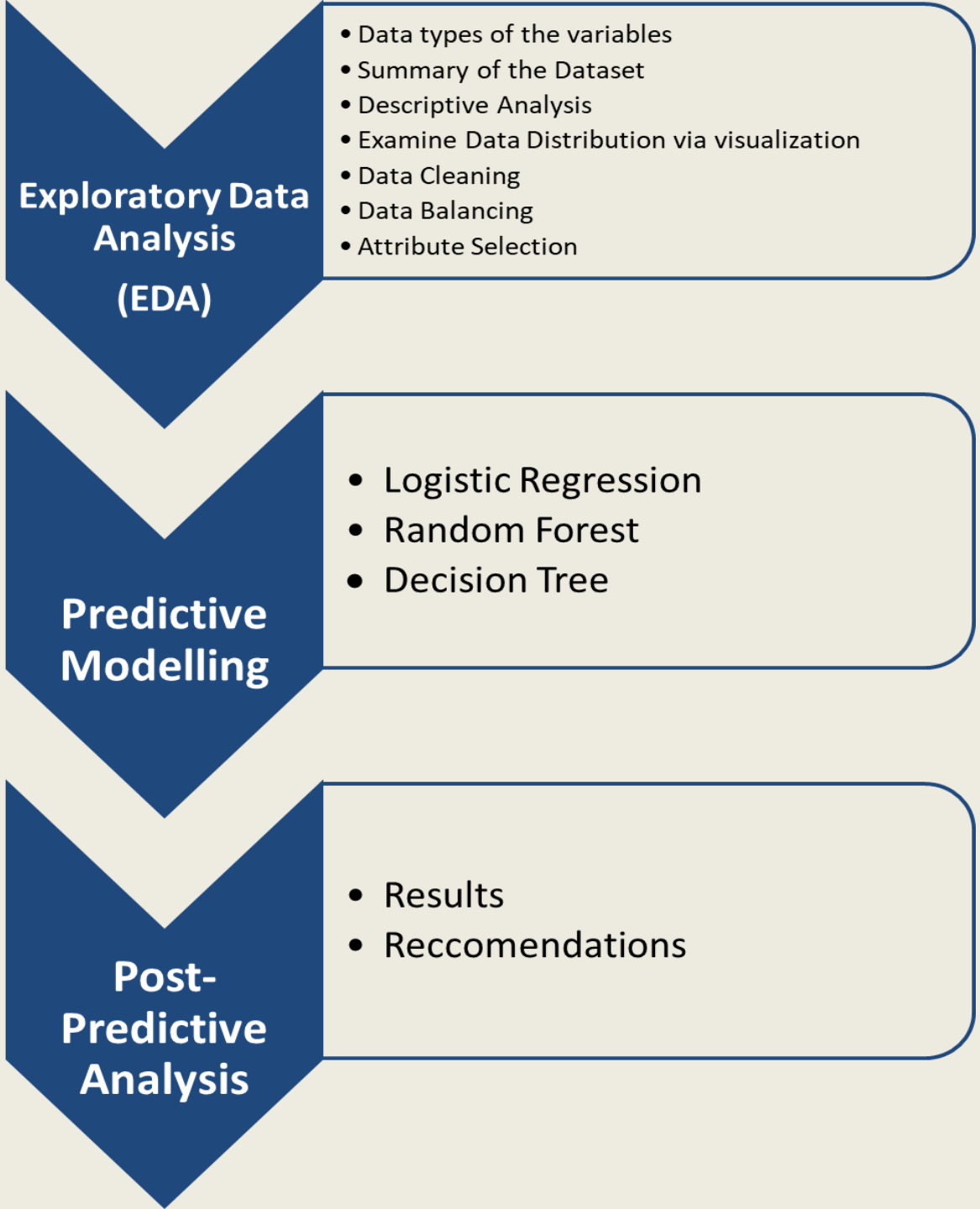
SEDEF.AKINLIKOCAK@RYERSON.CA

Approach

Step 1: Exploratory Data Analysis (EDA)

Step 2: Predictive Modelling

Step 3: Post-Predictive Analysis



Step 1: Exploratory Data Analysis (EDA)

- Data types of the Variables
- Summary of the Dataset
- Examine Data Distribution
- Density plots to see the Distribution of all the Variables
- Frequency table of all the columns that have character data types
- Total number of missing values in each column
- Data Cleaning: Character Variables
- Data Cleaning: Numeric Variables
- Density plots to see the Distribution of all the Variables after Cleaning
- Correlation Between all Numerical Variables
- Feature Selection

Step 2: Predictive Modelling

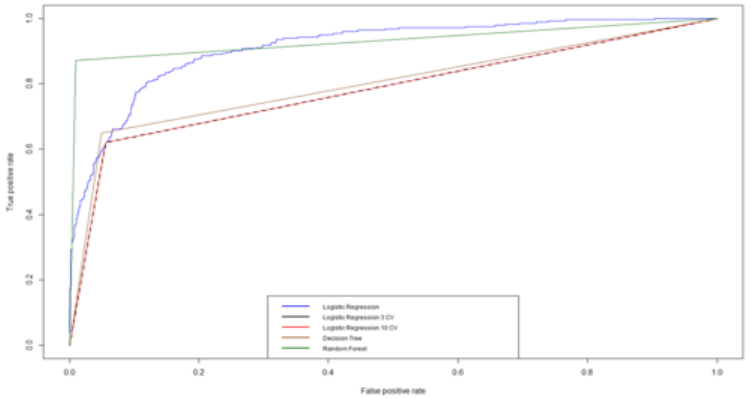
The 3 Machine Learning Models Used

- After reviewing multiple journal articles during the literature review these are the 3 models that were chosen:
  - Logistic Regression
  - Random Forest
  - Decision Tree

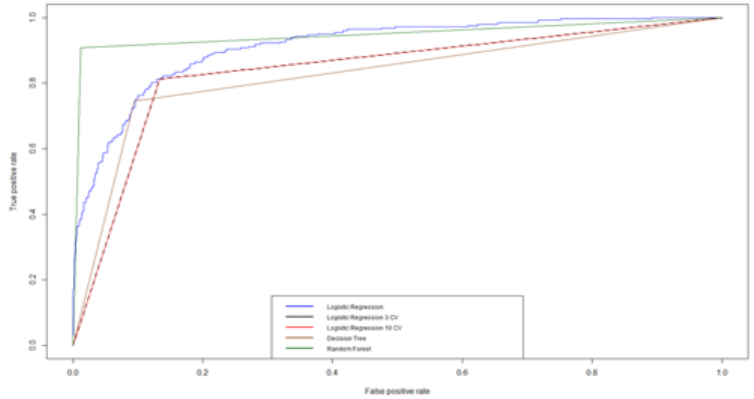
Step 3: Post-Predictive Analysis

Results

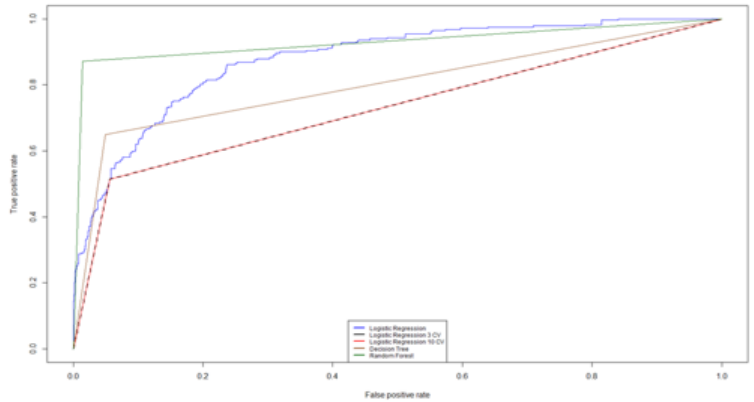
Unbalanced Training Set Before Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)² +(1-specificity)²) for ROC
Logistic Regression	0.91	0.93	0.62	0.94	0.88	0.91	
Logistic Regression 3 CV	0.78	0.93	0.62	0.94	0.88	0.91	
Logistic Regression 10 CV	0.78	0.93	0.62	0.94	0.88	0.91	
Random Forest	0.93	0.53	0.87	0.99	0.97	0.98	
Decision Tree	0.80	0.89	0.65	0.95	0.89	0.93	



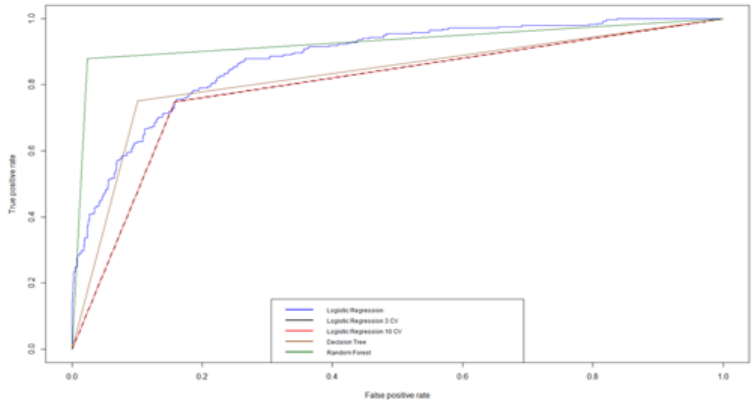
Balanced Training Set Before Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)²+ (1-specificity)²) for ROC
Logistic Regression	0.91	0.80	0.87	0.81	0.86	0.90	
Logistic Regression 3 CV	0.84	0.80	0.81	0.87	0.86	0.90	
Logistic Regression 10 CV	0.84	0.80	0.81	0.87	0.86	0.90	
Random Forest	0.95	0.46	0.91	0.99	0.97	0.98	
Decision Tree	0.82	0.84	0.74	0.91	0.87	0.92	



Unbalanced Training Set After Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)² +(1-specificity)²) for ROC
Logistic Regression	0.88	1.04	0.51	0.94	0.85	0.91	
Logistic Regression 3 CV	0.73	1.04	0.51	0.94	0.85	0.92	
Logistic Regression 10 CV	0.73	1.04	0.51	0.94	0.85	0.92	
Random Forest	0.93	0.53	0.87	0.99	0.96	0.98	
Decision Tree	0.80	0.89	0.65	0.95	0.89	0.93	



Balanced Training Set After Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)²+ (1-specificity)²) for ROC
Logistic Regression	0.88	0.91	0.84	0.75	0.82	0.88	
Logistic Regression 3 CV	0.80	0.91	0.75	0.84	0.82	0.88	
Logistic Regression 10 CV	0.80	0.91	0.75	0.84	0.82	0.88	
Random Forest	0.93	0.54	0.88	0.98	0.96	0.97	
Decision Tree	0.83	0.84	0.75	0.90	0.87	0.91	



Recommendations

- Continue to collect more data. This will help with the overfitting issue and improve the performance of all the models.
- If they still want to use the current predictions, the top 8 variables to keep in mind would be:
    1) Tenure
    2) OrderCount
    3) Complain
    4) NumberOfAddress
    5) CashbackAmount
    6) DaySinceLastOrder
    7) NumberOfDeviceRegistered
    8) WarehouseToHome

Limitations

- Relatively small dataset
- Due to time restrictions and the scope of this course:
  - The planned use of the XGBoost model was replaced with Decision Tree
  - Simple oversampling used instead of Safe-Level-SMOTE
  - Automatic feature selection method and Variable Importance from Machine Learning Algorithms, Recursive Feature Elimination (RFE) used instead of Step wise Forward and Backward Selection

Thank You!