Data Science Challenge Tasks

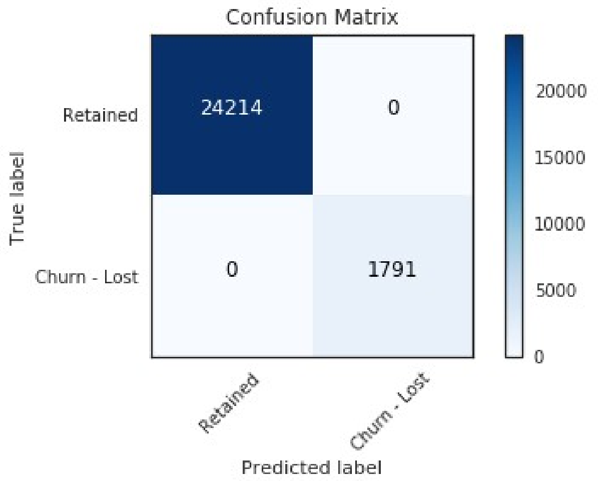
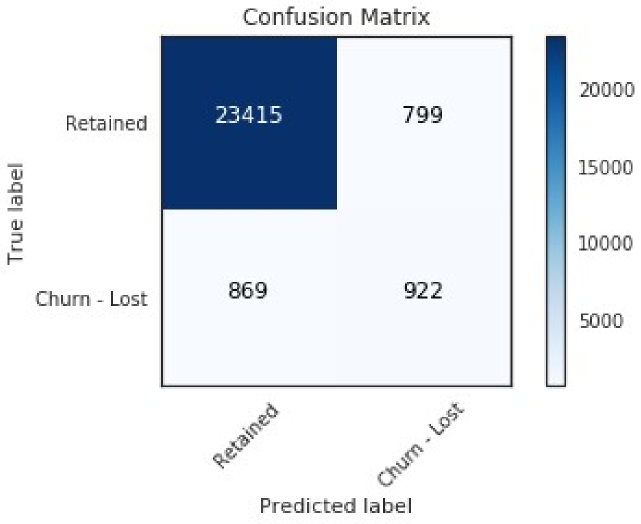
Task 1. Perform a descriptive analysis and visual exploration of key data elements (provide the code and output) and shortly report the (significance) of findings in plain language.
Answer :

Explorations Performed :
a. Numerical Features : min, mean, quartiles, max
b. Categorical Features : unique, top, freq
c. Binary : distribution with plot (few & hence easy to visualise)
d. Correlation : Vis (Double click to zoom)

Task 2. Build a simple model (decision tree or regression).
Answer : Completed. Please see section 2.2

1. Evaluation Metric : Confusion Matrix
2. Simple model : Decision Tree

Task 3 : Evaluate its performance and report findings.
Answer : With column "avg_online_days" included, the decision tree yields a perfect classification. Please see the matrix below for detailed report finding

With Column "avg_online_days"	Without Column "avg_online_days"
	

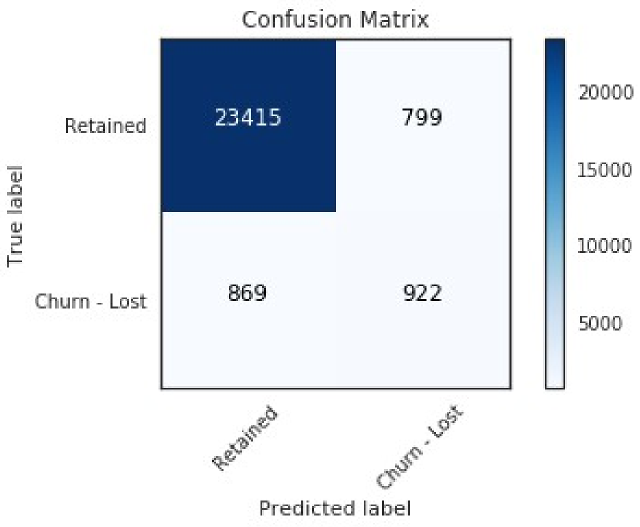
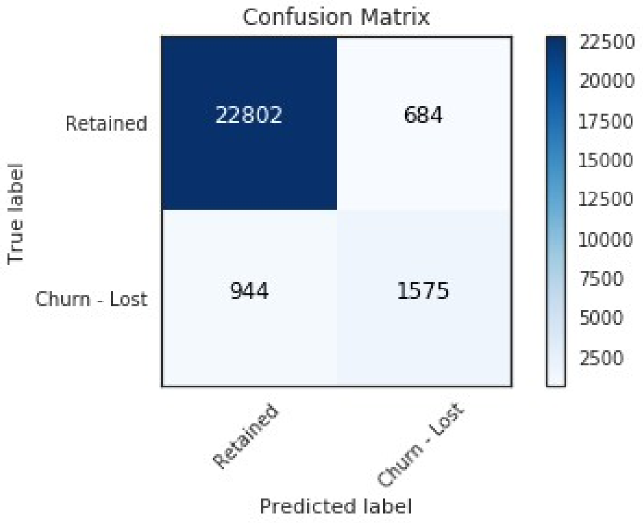
v1.2 : 1st task : If you have previous experience on lift analysis,
please perform a lift analysis on test data.

Task 4. Perform data cleansing, transformation, imputation on the data (do not exclude records with NA rather treat them).
Answer : Completed. See Section 4.0

Performed Actions

- Invalid Data - Negative Price : clipped to zero
- Invalid Data - Binary Feature : [1,2] to [1,0]
- Missing Values : Filled with median
- Uniform values features : Dropped
- Outliers : Preserved as is (Outliers are signals for outlier analysis)
- Feature varying Scale : Scaled
- Categorical Variables : Dummy variables created
- Correlation : PCA
- Biased Class Distribution : Oversampling Class 1 with SMOTE
- Sampling : Stratified

Task 5. Build second version of the model using advanced machine learning algorithms (random forest, neural network, deep learning, etc.) and evaluate it to compare the performance gain.
Answer : Built Random Forest Classifier & Neural Network

Simple Tree	Neural Network
	

Task 6. Based on available time, prioritize which steps would you further take to best optimize your model and then perform few of the most important steps to optimize the model. If you do not have sufficient time to cover this step, please provide a note on which steps would you take and why, what

method would you use for performing the task, and what is your expected outcome after performing those steps.

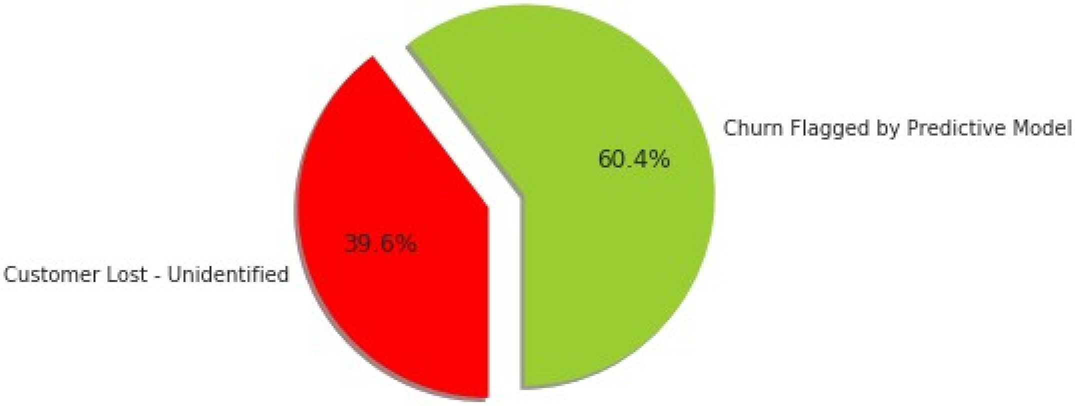
Answer :

- Data Skewness fix and continue - Improved performance as skewed data distribution can decrease performance.
- NN with regularisation : Regularisation penalises model complexity and hence overfitting and ensures is more generic to yield better results on. Possible cerrtain degree of overfitting may be present in the current model.
- Deep multiple layer neural network models - Increased performance because it can model much more complex feature space and boundaries.

Task 7. Prepare 2 slides to communicate the results to business.

Answer :

1. Our model can identify 60.4 % of the customers to be lost correctly, so that proper measure can be taken.


Churn Flagged by Predictive Model
Customer Lost - Unidentified