Churn Analysis

Zoha Afzal, zoha.afzal@ryerson.ca
Ali Nour, ali.nour@torontomu.ca
Semiu Kolapo, skolapo@torontomu.ca

RYERSON UNIVERSITY

Workload Distribution

Member Name	List of Tasks Performed
Ali Nour	- Predictive modeling/classification
Semiu Kolapo	- Data preparation - Baseline Models
Zoha Afzal	- Introduction - Conclusions and Recommendations

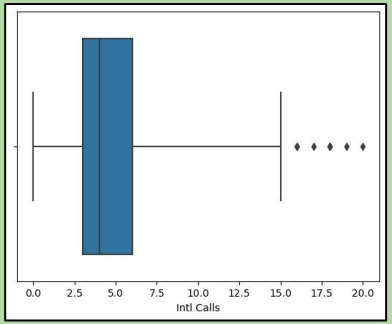
Introduction

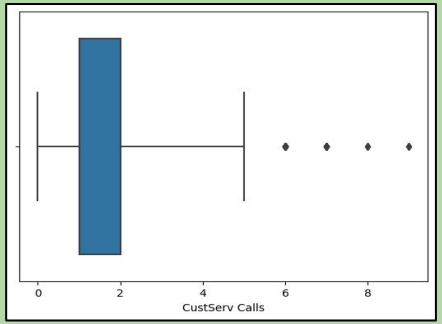
- **Problem**
  - Company experiencing customer churn, so it would like to know in advance which customers would churn in near future
- **Summary**
  - The “dataset of churn” is chosen to analyze
  - All of the rows are chosen except those that had extreme outliers in order to maintain maximum amount of data.
  - From 21 total attributes, of the attributes chosen 6 are qualitative and 15 are quantitative
- **Tools**
  - The tools used to analyze data was python
  - The predictive modelling was done through classification using decision tree and Naive Bayes.

Data Preparation

- **Step 1: Data Types/Missing Values**
  - Nominal/Qualitative, Quantitative
  - No Missing/Duplicate Values

- **Step 2: Removing Extreme Outliers** - 3 times the interquartile range (IQR)

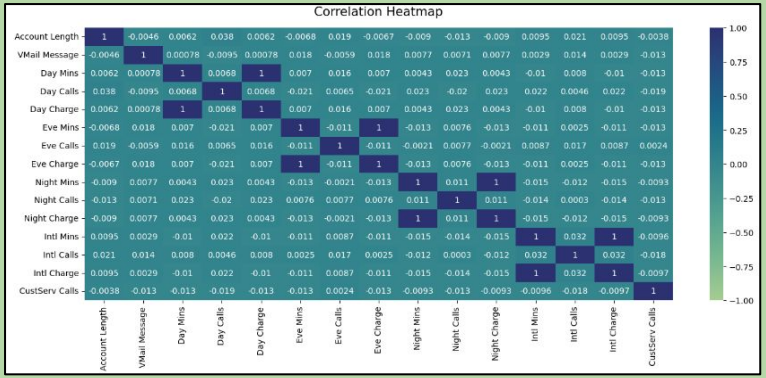




Data Preparation

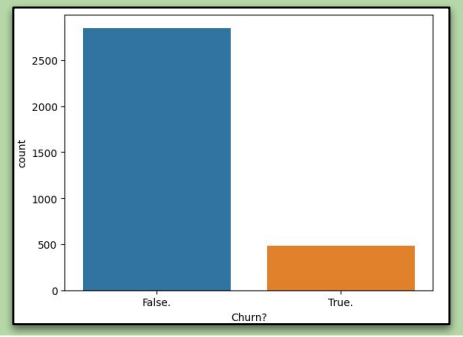
- **Step 3: Distribution Analysis** - Plotting histograms for all the numeric attributes
  - **Feature Scaling -** min max scaler

- **Step 4: Correlation Analysis**: The charge and minute attributes are correlated thus, minute attributes were removed from the dataset



Data Preparation

- **Step 5: Class Imbalance** - The RandomOverSampler method from the imblearn library was used to solve the imbalance distribution.
- **Step 6: Encode Categorical Data -** Label Encoder and One-Hot Encoder
  - Most models only accept numeric variables, preprocessing the categorical variables becomes a necessary step.



Predictive Modeling

Decision Tree

- We used simple train-test set split for the Dataset
- The Decision Tree analysis identifies the number of customer service calls as the most significant predictor for customer churn.
- The gini coefficient of 0.245 shows that this factor effectively distinguishes between churn and non-churn outcomes.
- The Decision Tree algorithm uses several attributes to make the decision, including the total number of voice mail messages, total day charge, total evening minutes, and international plan.

Predictive Modeling

Naïve Bayes

Confusion Matrix

- The model correctly predicted 758 true negatives and 66 true positives, but misclassified 92 false positives and 84 false negatives.
- Precision for the positive class is 0.42, recall is 0.44.
- Accuracy of the model is 0.82.

Performance Review

Naïve Bayes

Confusion Matrix

Performance Measures

Baseline

758	92
84	66

	Precision	Recall
FALSE	0.90	0.89
TRUE	0.42	0.44
Accuracy	**0.82**	

Selected features

316	534
34	116

	Precision	Recall
FALSE	0.90	0.37
TRUE	0.18	0.77
Accuracy	**0.43**	

RYERSON UNIVERSITY

Legend: TP, TN, FP, FN

Performance Review

Decision Tree

Baseline

Performance Measures

	Precision	Recall
FALSE	0.94	0.98
TRUE	0.83	0.64
Accuracy	**0.93**	

Selected features

	Precision	Recall
FALSE	0.89	1.00
TRUE	0.91	0.44
Accuracy	**0.89**	

Comparison

Baseline

Performs better in terms of accuracy (82%) and has a lower False Positive rate, indicating that it is better at correctly identifying negative cases.

Selected features

Has a higher True Positive rate, indicating that it is better at correctly identifying positive cases. The choice of algorithm depends on the specific requirements of the problem.

Therefore, the choice of the algorithm depends on the specific requirements of the problem. If the goal is to correctly identify positive cases, the algorithm with selected features may be more suitable, whereas if the goal is to correctly identify negative cases, the algorithm with all attributes may be more suitable.

Conclusion and Recommendation

Conclusion

- dataset ‘Churn?’ is completely imbalanced, with few extreme outliers but no missing or duplicate values
- distribution of the attributes mostly fall within normal range while three attributes are right-skewed and one that is left-skewed
- Decision Tree algorithm is more accurate with an accuracy rate of 0.89 compared to Naive Bayes model which has accuracy of 0.43

Recommendation

- improve number of customer service calls
- improve their international charge packages by correlating price and minutes