Churn Prediction Using IBM’s Telco Customer Churn Dataset

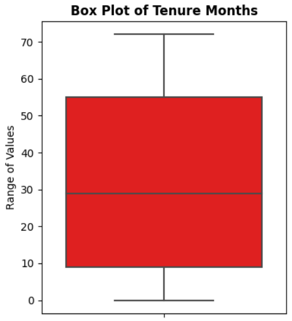
Data set and problem/task

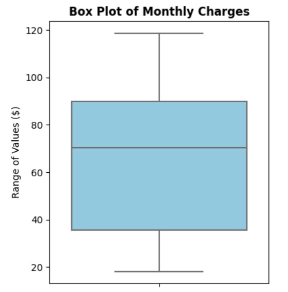
- The dataset contains info of **7043 customers** in CA.
- Used **19 features** out of the 32 available in the dataset. (LatLong feature was a noteworthy omission).
- The **class label** is the “**Churn Value**” column.
- There are **2 classes** in the dataset.
- **1** for the class label represents customer churned, & **0** represents customer was retained.
- Goal is to classify whether a customer will ***churn or*** be ***retained*** by the company.
- An example row from the dataset ============>
- The **Dummy** classifier from scikit-learn as the baseline, and the **KNN** (K=3) classifier were used.

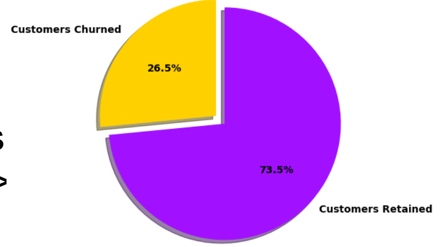
Gender	Female
Senior Citizen	No
Partner	Yes
Dependents	No
Tenure Months	65
Phone Service	Yes
Multiple Lines	No
Internet Service	DSL
Online Security	Yes
Online Backup	No
Device Protection	Yes
Tech Support	No
Streaming TV	No
Streaming Movies	No
Contract	Month-to-month
Paperless Billing	Yes
Payment Method	Mailed check
Monthly Charges	55.15
Total Charges	3673.15
Churn Value	0

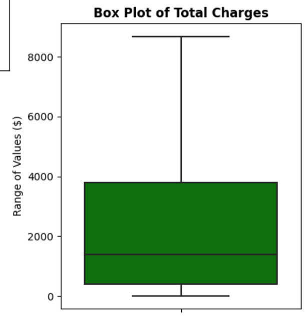
ML Challenges

- 1. **Encoding categorical data**: <======= In the dataset, 16 features contain categorical data.
- 2. **Feature scaling**: The 3 numerical features vary widely in scale & range, as can be seen on the right.
- 3. **Class imbalance**: Only 26.5% customers churned; significant class imbalance present ====>



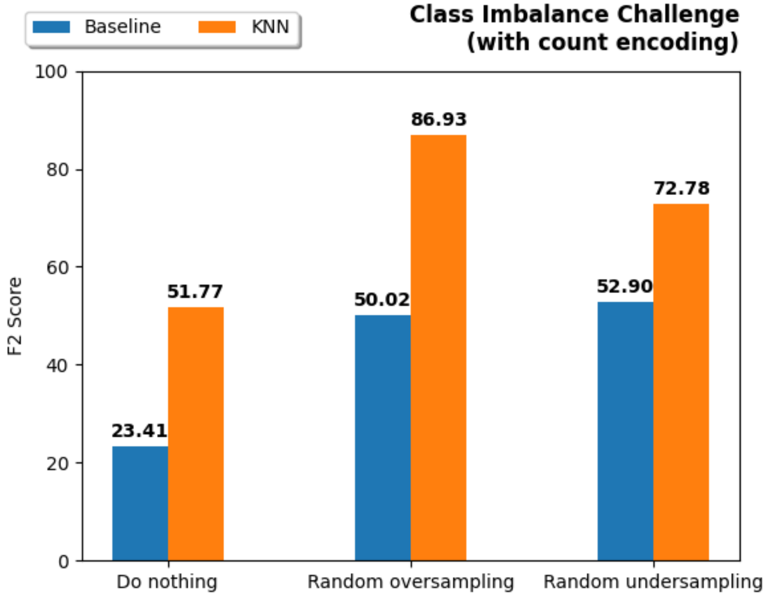






Key Experimental Result

- Tried 3 strategies to handle class imbalance:
  (i) **Do nothing**
  (ii) **Random oversampling**
  (iii) **Random undersampling**
- **F2** scores obtained using the Dummy & KNN classifiers for the 3 strategies ===========>
- F2 = **(5*P*R) / (4*P + R)**
  P = Precision
  R = Recall



AI 539 – Winter 2023

Future Work

The following points could be further investigated in future work:

- 1. *Most effective method for utilizing location-related information*:
  (i) Represent LatLong values as x,y,z co-ordinates (3 new features) & then scale them appropriately.
  (ii) Calculate the Haversine distance of a customer’s LatLong value from a major city and use it as a new feature.
- 2. Which *features* are the *most important* in predicting churn?
- 3. Use a *different classifier* (such as random forest) to compare its performance with KNN.

Shivam