TELECOM CHURN CASE STUDY – Insights and Recommendations

Vipin Suresh T (DS C65)
Shubham Kapur (DS C65)
Vishal Tyagi (DS C65)

Model Building Work Flow followed :

- Import required libraries, Read data, view data, info, describe
- Data preparation - removing redundant columns, handling missing values, filtering high value customers (target customer segment), tagging churners.
- Exploratory data analysis (EDA),
- Data imbalance handling by SMOTE oversampling
- Logistic regression, future selection using RFE and VIF check, Prediction and evaluation Metrics, Feature importance
- Decision tree , evaluation Metrics
- Random forests, evaluation Metrics, Feature importance
- Comparison of models
- Insights and recommendations models

Comparison for 3 models:

Evaluation Metrics	Logistic Regression	Decision Trees	Random Forests
Accuracy	0.82	0.89	0.93
Precision	0.29	0.4	0.55
Sensitivity (Recall)	0.79	0.71	0.68
Specificity	0.83	0.9	0.95
False Positive Rate	0.17	0.1	0.05
True Negative Prediction Rate	0.98	0.97	0.97
F-1 Score	0.42	0.51	0.61

1. Accuracy

- **Random Forests** shows the highest accuracy at 0.93, followed by **Decision Trees** at 0.89, and **Logistic Regression** at 0.82. For an imbalanced data, other metrics are also very important.

2. Precision

- **Random Forests** has the highest precision (0.55), which means it is better at correctly identifying actual churners (positive class) compared to the other models.
- **Decision Trees** follows with 0.40 precision, while **Logistic Regression** has the lowest precision at 0.29. Low precision in Logistic Regression means that a significant number of non-churners are being misclassified as churners.

3. Sensitivity (Recall)

- **Logistic Regression** performs the best in terms of recall (0.79), meaning it is good at identifying actual churners, even if it misclassifies some.
- **Decision Trees** and **Random Forests** have lower recall (0.71 and 0.68, respectively), meaning they miss a higher proportion of actual churners compared to Logistic Regression.

4. Specificity

- **Random Forests** has the highest specificity at 0.95, followed by **Decision Trees** (0.90) and **Logistic Regression** (0.83). High specificity means that Random Forests and Decision Trees are better at identifying non-churners.

5. False Positive Rate (FPR)

- **Random Forests** has the lowest FPR (0.05), which is a positive sign as it misclassifies fewer non-churners as churners.
- **Logistic Regression** has the highest FPR (0.17), indicating more false positives, which could be costly in business settings.

6. True Negative Prediction Rate

- All three models perform similarly in terms of predicting non-churners correctly, with **Logistic Regression** being marginally better (0.98).

7. F1-Score

- **Random Forests** has the highest F1-score at 0.61, followed by **Decision Trees** at 0.51, and **Logistic Regression** at 0.42.
    - The F1 score balances precision and recall, making **Random Forests** the most balanced model overall for churn prediction, with better precision and a reasonable recall.

Summary

- **Random Forests** performs the best overall, with the highest accuracy, precision, and F1-score. It is more reliable in minimizing false positives while also correctly identifying non-churners.
- **Logistic Regression** performs well in terms of recall (sensitivity), making it better at identifying actual churners but at the cost of low precision and higher false positives.
- **Decision Trees** is a middle-ground model, with fairly balanced metrics, but it is outperformed by Random Forests in almost every category.
- If identifying all possible churners is the priority (to avoid false negatives), **Logistic Regression** may be more useful.
- However, if we are looking for a more balanced model with good performance across both precision and recall, **Random Forests** is likely the best choice for your telecom churn prediction case.

Insights

1. The **Top features** which indicate that the customer is going to churn are -
    1. Local incoming and outgoing MOU (to mobile and fixed line) during action phase.
    2. Roaming incoming and Outgoing MOU
    3. Special incoming MOU during action phase.
    4. Average revenue per user.
    5. Last day recharge amount.
    6. STD incoming MOU (other operator)
    7. Offnet MOU and Onnet MOU
    8. Night pack user or not
    9. Max recharge amount for calls and data.
    10. Volume of 2G and 3G usage.

2. **Age on Network** - Churn possibility is higher for users with less than 12 months. High value users with less than 12 months on network need to be carefully monitored. Special offers can be roilled out which will motivate the user to move on to second year.
3. **Average revenue per user(ARPU)** : A sudden sharp reduction in ARPU in subsequent months is a clear indicator of churn. Customers to be contacted and steps to be taken to retain the customer when such a behaviour is observed.
4. Similerly a suddent reduction in Minutes of usage **(MOU)-** calls within same network, calls outside network, local, incoming, outgoing, roaming - are also strong indicators of possibility to churn.
5. Similarly - Sudden reduction **Recharge amount** for talk time and data also indicate possibility to churn.
6. Sudden reduction in the usage of **3g data services** and usage of services with validity less than a month (for example 3g Sachets) indicate possibility to churn.
7. Users who are using more **Roaminng** in Outgoing and Incoming calls, are very likely to churn. Compnay can focus on them to retain them.