Table of Content

Telecom Churn Prediction Analysis Report

1. Findings:

Data Overview:

The Churn Dataset consists of cleaned customer activity data and churn labels. Two datasets are provided:

- Churn-bigml-80: This dataset, comprising a larger portion of the data, is split by an 80/20 ratio. It serves as the primary dataset for training and cross-validation purposes.
- Churn-bigml-20: This smaller subset, also split from the same batch, follows the same features and churn labels. It is reserved for final testing and evaluating model performance.
-

Utilising Churn-80 for training and cross-validation ensures a more robust model development process, while Churn-20 allows for unbiased evaluation of the final model's performance.

Churn Rate:

Within the Churn-80 dataset, the churn rate is calculated to be 14.6%, providing insight into the prevalence of churn within the training and cross-validation dataset.

Feature Analysis:

Customer Service Calls Analysis:

- Investigated the relationship between the number of customer service calls and the likelihood of churn.
- Found that as the number of customer service calls increases, there is a corresponding increase in the churn rate.
- Suggests that customers who contact customer service more frequently are more likely to churn, possibly indicating dissatisfaction with the service or unresolved issues.

Churn Imbalance:

- Identified a significant class imbalance within the dataset.
- Disproportionately higher number of non-churned customers compared to churned customers.

- Imbalance may pose challenges during model training and evaluation.
- Plan to explore techniques such as oversampling, undersampling, or adjusting class weights during model training to address the issue.

International Plan Analysis:

- Analysed churn rates among customers with and without international plans.
- Found that approximately half of the customers with international plans churn, compared to a lower churn rate among customers without international plans.
- Suggests that the presence of an international plan may play a significant role in predicting churn.
- Customers with international plans may have distinct behaviour patterns or service expectations influencing their likelihood of churning.

Feature Engineering:

Handling Categorical Features:

State:

Encode the "State" feature using ordinal encoding to represent the states with integer labels based on their ordinal relationships. This preserves the ordinality of the states in the dataset.

International Plan and VoiceMail Plan:

Encode the "International Plan" and "Voice Mail Plan" features using ordinal encoding if there is an inherent order to the categories. For instance, if there is an implicit order such as "Yes" being considered higher than "No", ordinal encoding can be applied to reflect this relationship.

Churn:

Encode the "Churn" feature using label encoding to convert the boolean values into numeric labels. Assign "False" the label 0 and "True" the label 1, representing non-churned and churned customers, respectively.

Creating New Features:

Generate a new feature called "Total Charge" by summing up the charges for day, evening, night, and international calls. This feature is expected to have a significant

relationship with churn, as higher total charges may indicate higher usage or dissatisfaction with the service, leading to churn.

Explored other feature engineering techniques such as binning and creating features like "Total Calls," but found that they had a negative impact on model performance.

After experimentation, it was determined that "Total Charge" had the most significant positive effect on predicting churn, outperforming other engineered features. Also proved from the Decision tree Feature importance.

Scaling Numerical Features:

Scale numerical features using Robust Scaler to robustly handle outliers. Robust Scaler scales the features using statistics that are robust to outliers, such as the median and interquartile range (IQR).

This ensures that the presence of outliers does not disproportionately influence the scaling process and helps improve the performance of the models.

Handling Imbalanced Data:

Experimented with various techniques to address class imbalance, including oversampling the minority class (churned customers) and undersampling the majority class (non-churned customers).

Found that the best results were achieved with BorderlineSMOTE, an oversampling technique specifically designed to address imbalanced datasets.

BorderlineSMOTE generates synthetic samples near the decision boundary between classes, helping to improve the generalisation ability of the models while avoiding overfitting.

This adjustment includes the use of BorderlineSMOTE for handling the imbalanced data, which resulted in the best performance for the churn prediction models.

Model Evaluation:

The performance metrics for various churn prediction models are as follows:

Accuracy:

Measures the overall correctness of the model's predictions, calculated as the ratio of correctly predicted instances to the total number of instances.

Precision:

Indicates the proportion of true positive predictions among all positive predictions made by the model, highlighting the model's ability to minimise false positives.

Recall (Sensitivity):

Measures the proportion of true positive predictions that were correctly identified by the model among all actual positive instances, indicating the model's ability to capture all positive instances.

F1-Score:

Harmonic mean of precision and recall, providing a balanced measure of the model's accuracy in terms of both false positives and false negatives.
These metrics will help evaluate the effectiveness of each churn prediction model in accurately identifying churned customers while minimising false predictions.

2. Model Performance Comparison:

Linear Models:

Linear Regression Model (NF):

- Precision: 0.3469387755102041
- Recall: 0.6455696202531646
- F1 Score: 0.45132743362831856
- Accuracy: 0.7677902621722846

Linear Regression Model (F):

- Precision: 0.8848920863309353
- Recall: 0.8870192307692307
- F1 Score: 0.885954381752701
- Accuracy: 0.8958333333333334

Gradient Regression Model (NF):

- Precision: 0.362962962962963

- Recall: 0.620253164556962
- F1 Score: 0.45794392523364486
- Accuracy: 0.7827715355805244

Gradient Regression Model (F):

- Precision: 0.6745562130177515
- Recall: 0.8221153846153846
- F1 Score: 0.7410617551462622
- Accuracy: 0.7379385964912281

Logistic Models:

Logistic Regression Model (NF):

- Precision: 0.3548387096774194
- Recall: 0.6962025316455697
- F1 Score: 0.4700854700854701
- Accuracy: 0.7677902621722846

Logistic Regression Model (F):

- Precision: 0.8758620689655172
- Recall: 0.9158653846153846
- F1 Score: 0.8954171562867215
- Accuracy: 0.9024122807017544

SGD Regression Model (NF):

- Precision: 0.4418604651162791
- Recall: 0.4810126582278481
- F1 Score: 0.46060606060606063
- Accuracy: 0.8333333333333334

SGD Regression Model (F):

- Precision: 0.8994708994708994
- Recall: 0.8173076923076923
- F1 Score: 0.8564231738035264
- Accuracy: 0.875

Decision Tree:

Decision Tree Model (NF):

- Precision: 0.684931506849315

- Recall: 0.6329113924050633
- F1 Score: 0.6578947368421053
- Accuracy: 0.9026217228464419

Decision Tree Model (F):

- Precision: 0.9070294784580499
- Recall: 0.9615384615384616
- F1 Score: 0.9334889148191365
- Accuracy: 0.9375

Neural Network:

Neural Network Model (NF):

- Precision: 0.8051948051948052
- Recall: 0.6526315789473685
- F1 Score: 0.7209302325581395
- Accuracy: 0.9280359820089955

Neural Network Model (F):

- Precision: 0.6822429906542056
- Recall: 0.7684210526315789
- F1 Score: 0.7227722772277227
- Accuracy: 0.9160419790104948

XGBoost:

XGBoost Model (NF):

- Precision: 0.9375
- Recall: 0.7894736842105263
- F1 Score: 0.8571428571428571
- Accuracy: 0.9625187406296851

XGBoost Model (F):

- Precision: 0.9540229885057471
- Recall: 0.8736842105263158
- F1 Score: 0.9120879120879121
- Accuracy: 0.9760119940029985

Random Forest:

Random Forest Model (NF):

- Precision: 0.9324324324324325

- Recall: 0.7263157894736842
- F1 Score: 0.8165680473372781
- Accuracy: 0.9535232383808095

Random Forest Model (F):

- Precision: 0.9761904761904762
- Recall: 0.8631578947368421
- F1 Score: 0.9162011173184358
- Accuracy: 0.9775112443778111

Comparison Table

Model Name	Type	Accuracy	Precision	Recall	F1-Score
Linear Regression Model (NF)	Linear	0.76	0.34	0.64	0.45
Linear Regression Model (F)	Linear	0.89	0.88	0.88	0.88
Gradient Regression Model (NF)	Linear	0.78	0.36	0.62	0.45
Gradient Regression Model (F)	Linear	0.73	0.67	0.82	0.74
Logistic Regression Model (NF)	Logistic	0.76	0.35	0.69	0.47
Logistic Regression Model (F)	Logistic	0.90	0.87	0.91	0.89
SGD Regression Model (NF)	Logistic	0.83	0.44	0.48	0.46

Model Name	Type	Accuracy	Precision	Recall	F1-Score
SGD Regression Model (F)	Logistic	0.87	0.89	0.81	0.85
Decision Tree Model (NF)	Tree	0.90	0.68	0.63	0.65
Decision Tree Model (F)	Tree	0.93	0.90	0.96	0.93
Neural Network Model (NF)	Neural Network	0.92	0.80	0.65	0.72
Neural Network Model (F)	Neural Network	0.91	0.68	0.76	0.72
XGBoost Model (NF)	Ensemble	0.96	0.93	0.78	0.85
XGBoost Model (F)	Ensemble	0.97	0.95	0.87	0.91
Random Forest Model (NF)	Ensemble	0.95	0.93	0.72	0.81
Random Forest Model (F)	Ensemble	0.97	0.97	0.86	0.91

3. Suggestions for Improvement:

Data Augmentation:

Expand the dataset by gathering more customer activity data, including demographics, usage patterns, and interactions with customer support. Incorporating user feedback data obtained through NLP techniques, such as sentiment analysis of call reviews, can provide valuable insights into customer sentiments and preferences.

Geolocation Analysis:

Integrate geolocation data to identify regions with weak signal coverage or network issues. Addressing infrastructure problems in these areas can help mitigate churn risk and improve service quality.

Internet Usage Factor:

Include Features related to internet plan, internet usage and browsing behaviour to gain a comprehensive understanding of customer interactions with various services. Analysing internet usage patterns, data consumption, and application usage can reveal hidden factors contributing to churn.

Price Optimization:

Analyse the relationship between pricing changes and churn rates to inform pricing strategies. Offering competitive pricing plans or personalised discounts, particularly for international plans, may incentivize customers to stay and reduce churn.

NLP and Sentiment Analysis:

Implement a system similar to ABC Messenger to collect customer reviews after calls. Apply NLP techniques to extract and analyse these reviews, conducting sentiment analysis to understand customer satisfaction levels and identify patterns or recurring issues. This insight can be used to improve service quality, customer support, and overall customer experience, ultimately reducing churn rates.

4. Conclusion:

After evaluating various machine learning models for churn prediction, it is evident that ensemble models, particularly those based on decision trees, demonstrate the highest efficiency and effectiveness in predicting customer churn. Ensemble models, such as Random Forest and XGBoost, leverage the strength of multiple weak learners to improve predictive accuracy and generalisation performance.

Through rigorous experimentation and analysis, it was observed that ensemble models consistently outperformed other individual models across multiple evaluation metrics, including precision, recall, F1 score, and accuracy. These models demonstrated robustness in handling complex patterns within the dataset and exhibited superior predictive capabilities in identifying potential churners..

In conclusion, the utilisation of ensemble models, particularly decision tree-based approaches, represents a promising avenue for enhancing churn prediction accuracy and maximising customer retention efforts in the telecommunications industry.