Custumer Churn Prediction

Predicting Customer Churn with Regression-Based and Tree-Based Methods

In this data modelling report, the data modeler utilized Orange Telecom Churn Dataset to build and evaluate machine learning models that predict whether a customer will churn.

DSC 1107 CAPSTONE PROJECT

BY ROMAND LANSANGAN

Objective

Making Sense of Raw Data

To gain the knowledge as to why churners decide to leave is to gain key information to develop a better business model for the company. A better business model that not only retain its clients but also a business model that could be used to entice the market.

	dtypes	nunique	unique
State	object	51	[KS, OH, NJ, OK, AL, MA, MO, WV, RI, IA, MT, I...
Account length	int64	205	[128, 107, 137, 84, 75, 118, 121, 147, 141, 74...
Area code	int64	3	[415, 408, 510]
International plan	object	2	[No, Yes]
Voice mail plan	object	2	[Yes, No]
Number vmail messages	int64	42	[25, 26, 0, 24, 37, 27, 33, 39, 41, 28, 30, 34...
Total day minutes	float64	1489	[265.1, 161.6, 243.4, 299.4, 166.7, 223.4, 218...
Total day calls	int64	115	[110, 123, 114, 71, 113, 98, 88, 79, 84, 127, ...
Total day charge	float64	1489	[45.07, 27.47, 41.38, 50.9, 28.34, 37.98, 37.0...
Total eve minutes	float64	1442	[197.4, 195.5, 121.2, 61.9, 148.3, 220.6, 348....
Total eve calls	int64	120	[99, 103, 110, 88, 122, 101, 108, 94, 111, 148...
Total eve charge	float64	1301	[16.78, 16.62, 10.3, 5.26, 12.61, 18.75, 29.62...
Total night minutes	float64	1444	[244.7, 254.4, 162.6, 196.9, 186.9, 203.9, 212...
Total night calls	int64	118	[91, 103, 104, 89, 121, 118, 96, 97, 94, 128, ...
Total night charge	float64	885	[11.01, 11.45, 7.32, 8.86, 8.41, 9.18, 9.57, 9...
Total intl minutes	float64	158	[10.0, 13.7, 12.2, 6.6, 10.1, 6.3, 7.5, 7.1, 1...
Total intl calls	int64	21	[3, 5, 7, 6, 2, 4, 19, 10, 9, 15, 8, 1, 11, 0,...
Total intl charge	float64	158	[2.7, 3.7, 3.29, 1.78, 2.73, 1.7, 2.03, 1.92, ...
Customer service calls	int64	10	[1, 0, 2, 3, 4, 5, 7, 9, 6, 8]
Churn	bool	2	[False, True]

Project Overview

5 Key Steps to Analyze Data

In this data modelling report, the data modeler utilized Orange Telecom Churn Dataset to build and evaluate machine learning models that predict whether a customer will churn. The modeler used both regression-based (particularly Logistic Regression) and tree-based (particularly Random Forest) to attain the objectives. The modeler also provided a comparison in the performance of both models and some nuance consideration.

1. EXPLORATORY DATA ANALYSIS
2. MODELING
3. MODEL EVALUATION
4. RESULTS AND RECOMMENDATIONS
5. CONCLUSION

Exploratory Data Analysis

Significant Imbalance

Both Figure 1 and Figure 2 show the imbalance between churners and non-churners. It becomes clear that a strict resampling technique was imperative

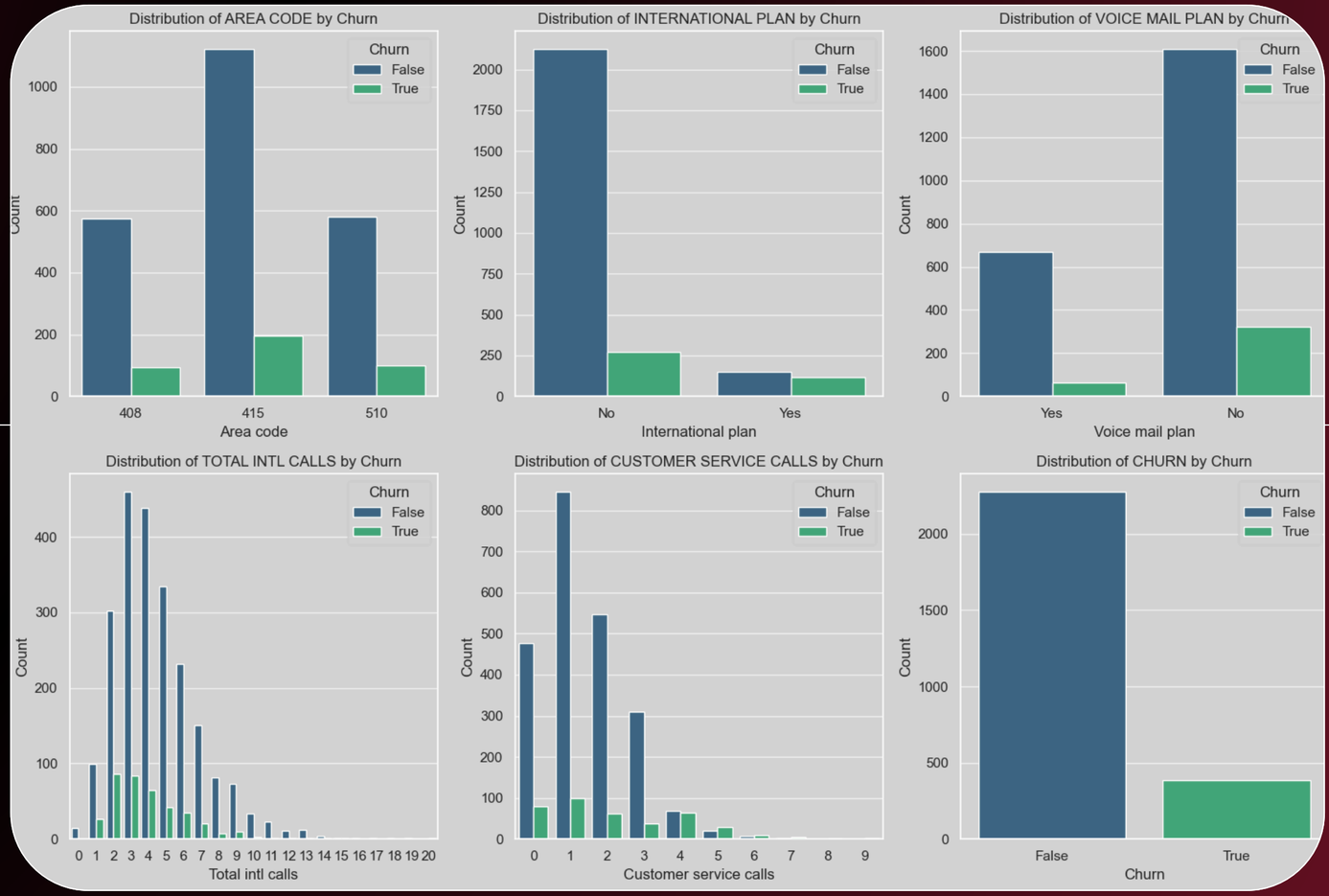


Figure 1: Distribution plots segmented by churn status for the categorical variables, showing differences between customers who churned (True) versus those who stayed (False).

Exploratory Data Analysis

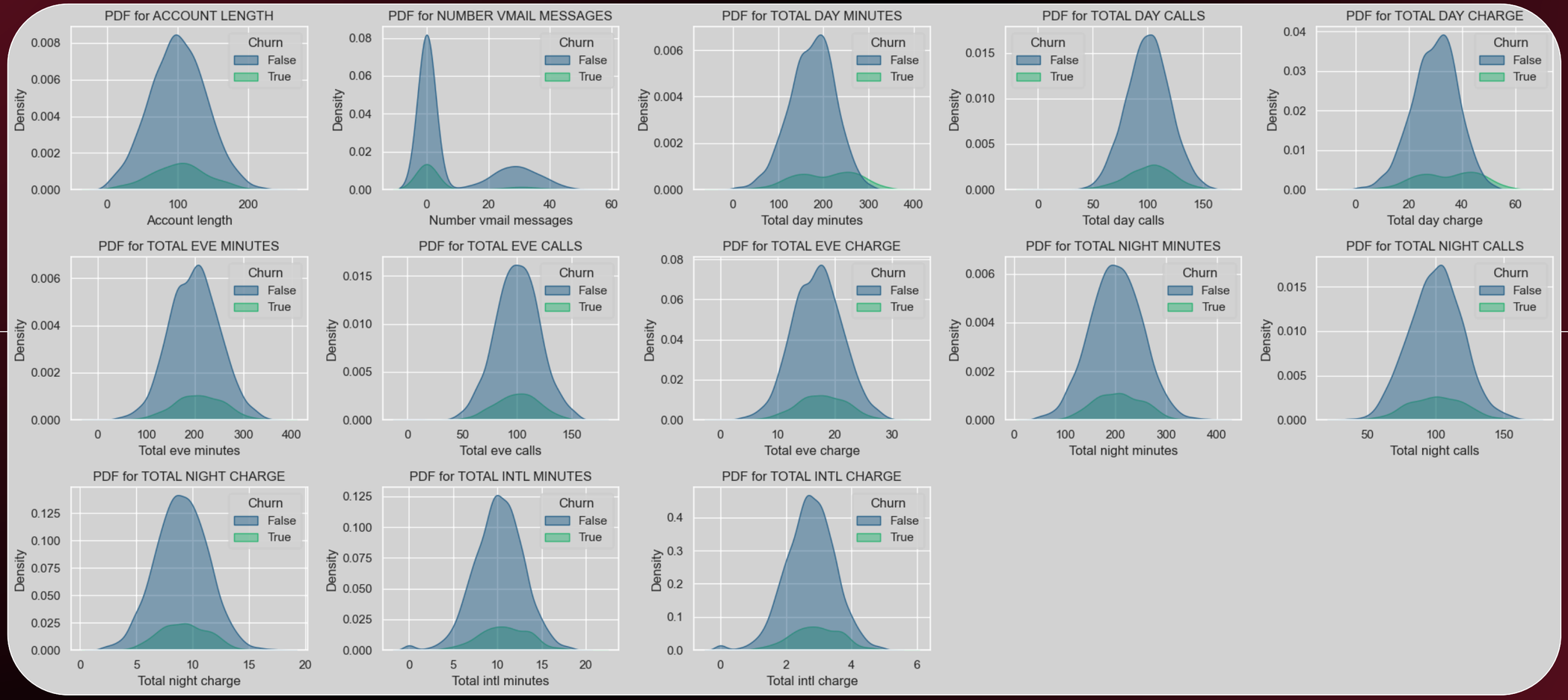


Figure 2: Probability density function (PDF) of continuous variables segmented by churn status, highlighting differences in usage patterns between customers who churned versus those who remained.

Exploratory Data Analysis

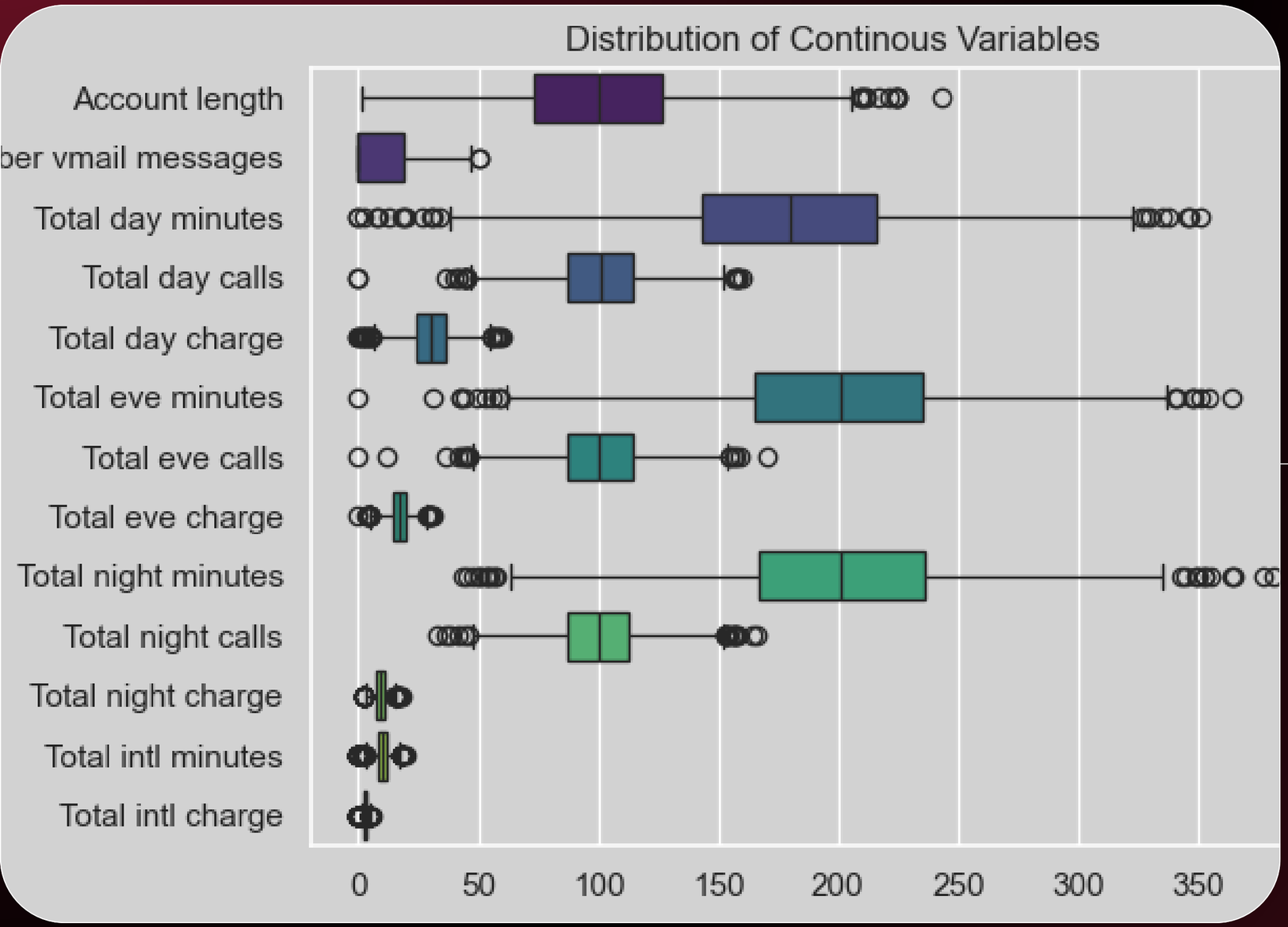


Figure 3: Box plot showing the distribution of continuous telecommunications variables, with metrics like account length and calling patterns (minutes, calls, charges) across day, evening, night, and international categories.

Outliers

Account length	0.004501
Number vmail messages	0.000750
Total day minutes	0.007877
Total day calls	0.006752
Total day charge	0.007877
Total eve minutes	0.006377
Total eve calls	0.005626
Total eve charge	0.006377
Total night minutes	0.008252
Total night calls	0.007127
Total night charge	0.008252
Total intl minutes	0.013878
Total intl charge	0.015004

Table 2: Proportion of outliers

The outliers in “Total International charge” and “Total International Minutes” were removed (40 rows). The rest were retained due to the insignificance of the frequency and to the fact that they might contain some key information for the model to be robust to outliers.

Exploratory Data Analysis

High Correlation

Some “total minutes” and their corresponding “total charges” are perfectly correlated. It means that it is safe to drop any of the pair. The modeler decided to drop the “total charges” columns.

The modeler decided to drop the “Voice mail plan” since its correlation to the response variable “Churn” is lower (-0.09) than the that of “Number vmail messages” (-0.10).

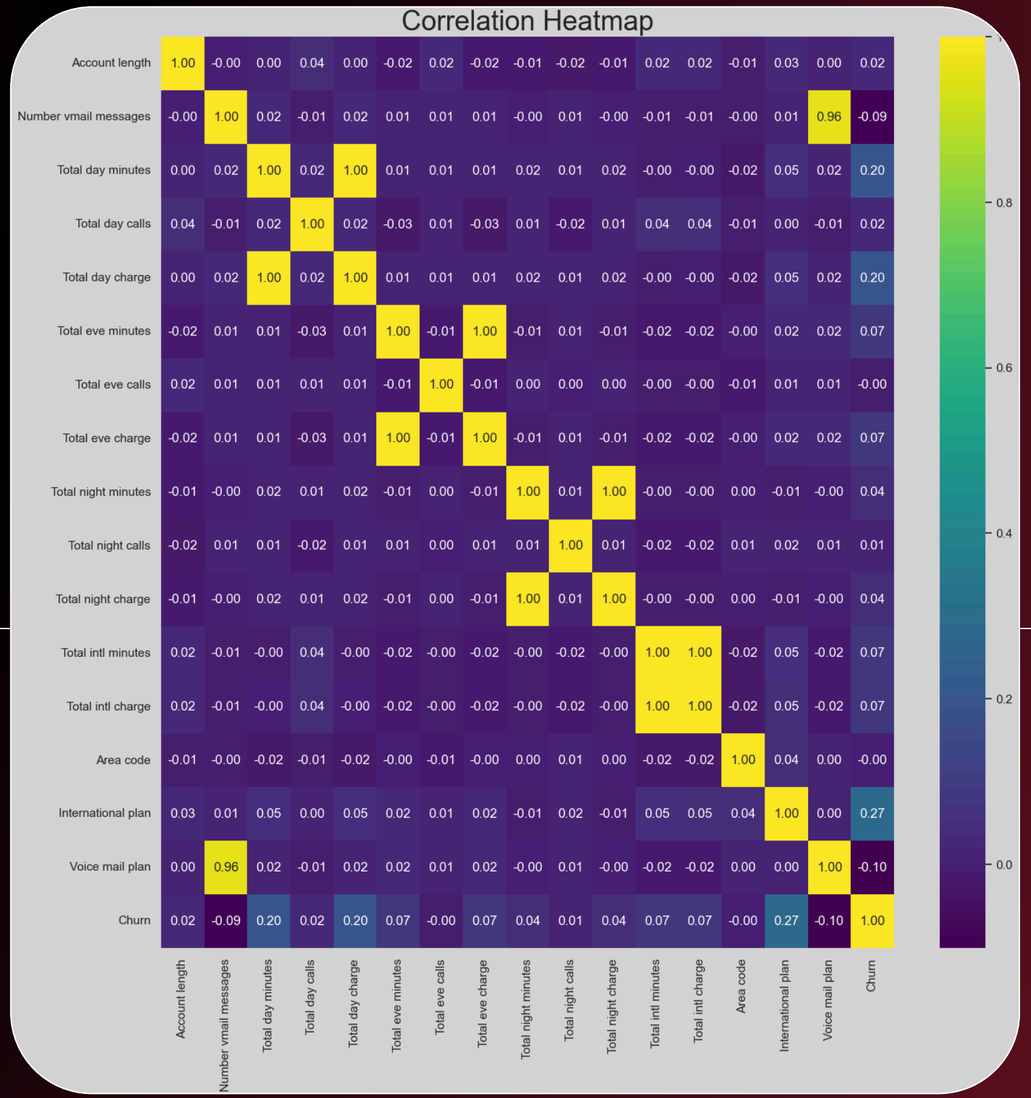


Figure 4: Correlation heatmap of telecommunications variables

Exploratory Data Analysis

Wrangle Function

1. INPUT: DATA AND OPTIONAL STANDARDSCALER
2. MODE DETECTION: TEST MODE IF STANDARDSCALER PROVIDED (NO OUTLIER REMOVAL), TRAIN MODE OTHERWISE (OUTLIER REMOVAL + NEW STANDARDSCALER)
3. PROCESS CATEGORICAL VARIABLES PER SECTION 2.4.1
4. PROCESS CONTINUOUS VARIABLES PER SECTION 2.4.2
5. CONVERT BINARY VARIABLES TO 0/1
6. DROP SPECIFIED VARIABLES: CUSTOMER SERVICE CALLS, TOTAL INTL CALLS, VOICE MAIL PLAN, TOTAL DAY CHARGE, TOTAL EVE CHARGE, TOTAL NIGHT CHARGE
7. RETURN CLEANED DATA AND STANDARDSCALER

Rows decreased to 2666 in the training set (no removal for test set) and number of columns increased to 27.

Modeling

Baseline normalized frequency of the majority class

Reseampling Synthetic Minority Over-sampling Technique (SMOTE) which is an oversampling with awareness of its neighbors and Edited Nearest Neighbors (ENN) which is an under sampling technique that remove some noise and clean the boundaries between classes.

Hyperparameter Tuning Result

Logistic Regression	Random Forest
{'C': 1, 'penalty': 'l2', 'solver': 'newton-cg'}	{'bootstrap': True, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

Table 3: Best parameters for each model after hyperparameter tuning.

Model Evaluation

Random Forest model upped the logistic regression by a wider margin for all metrics: accuracy, precision, recall, and f1-score.

	model	resampled_train	unsampled_train	test_set	test_set_log_loss
0	baseline	0.367982	0.853770	0.857571	5.133654
1	logistic regression	0.894328	0.790175	0.755622	0.599926
2	random forest	1.000000	0.905560	0.820090	0.419354

Table 4: Model performance comparison showing that while baseline performed the best at test set (85.76%; which is to be expected given the imbalance). The random forest achieves the highest accuracy with 100% in resampled train, 90.56% on unsampled train, 82.01% on test and the lowest log loss (0.419), outperforming logistic regression and baseline models.

Logitic Regression Model:

	precision	recall	f1-score	support
0	0.94	0.76	0.84	572
1	0.33	0.71	0.45	95
accuracy			0.76	667
macro avg	0.64	0.73	0.65	667
weighted avg	0.85	0.76	0.79	667

Table 5.1: Logistic Regression classification metrics

Random Forest Model:

	precision	recall	f1-score	support
0	0.97	0.82	0.89	572
1	0.43	0.82	0.57	95
accuracy			0.82	667
macro avg	0.70	0.82	0.73	667
weighted avg	0.89	0.82	0.84	667

Table 5.2: Random Forest classification metrics

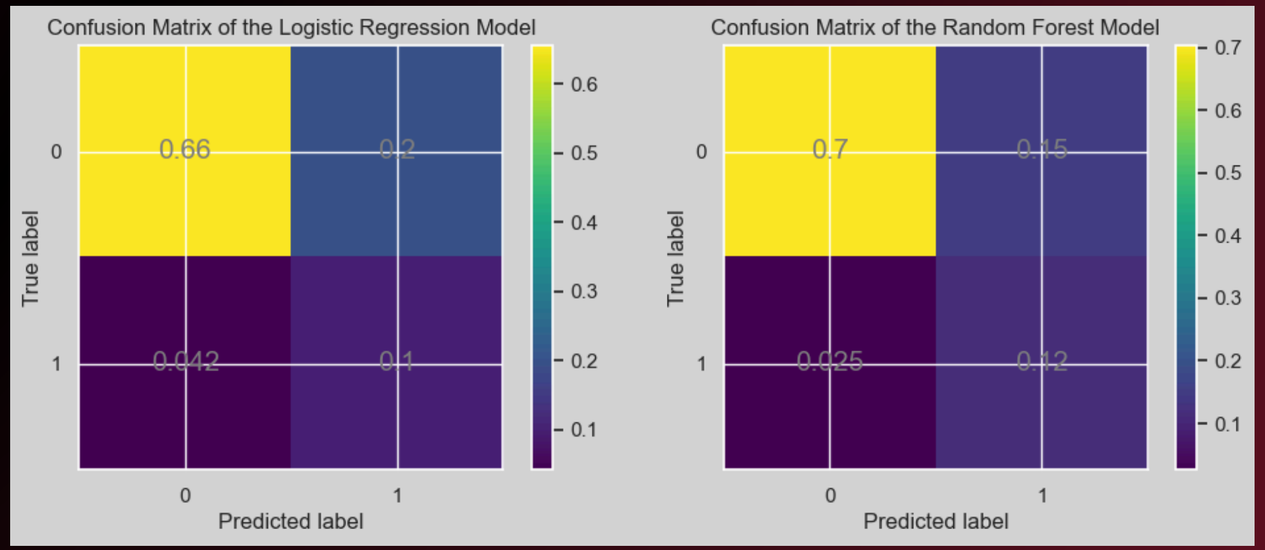


Figure 5: Confusion matrices comparing model performance. Random Forest (right) shows better overall accuracy with 70% true negatives and 12% true positives, while Logistic Regression (left) shows 66% true negatives and 10% true positives.

Feature Importance

Stark Difference Between Model

Aside from one outlier, **Random Forest** resulted to a more balanced coefficients. This might indicate that the model was able to catch some non-linear relationship between the predictors and response.

On the other hand, **Logistic Regression** offers more distinct importance. This is a point because it is easier to interpret and one can clearly ascertain which features are more important than the others.

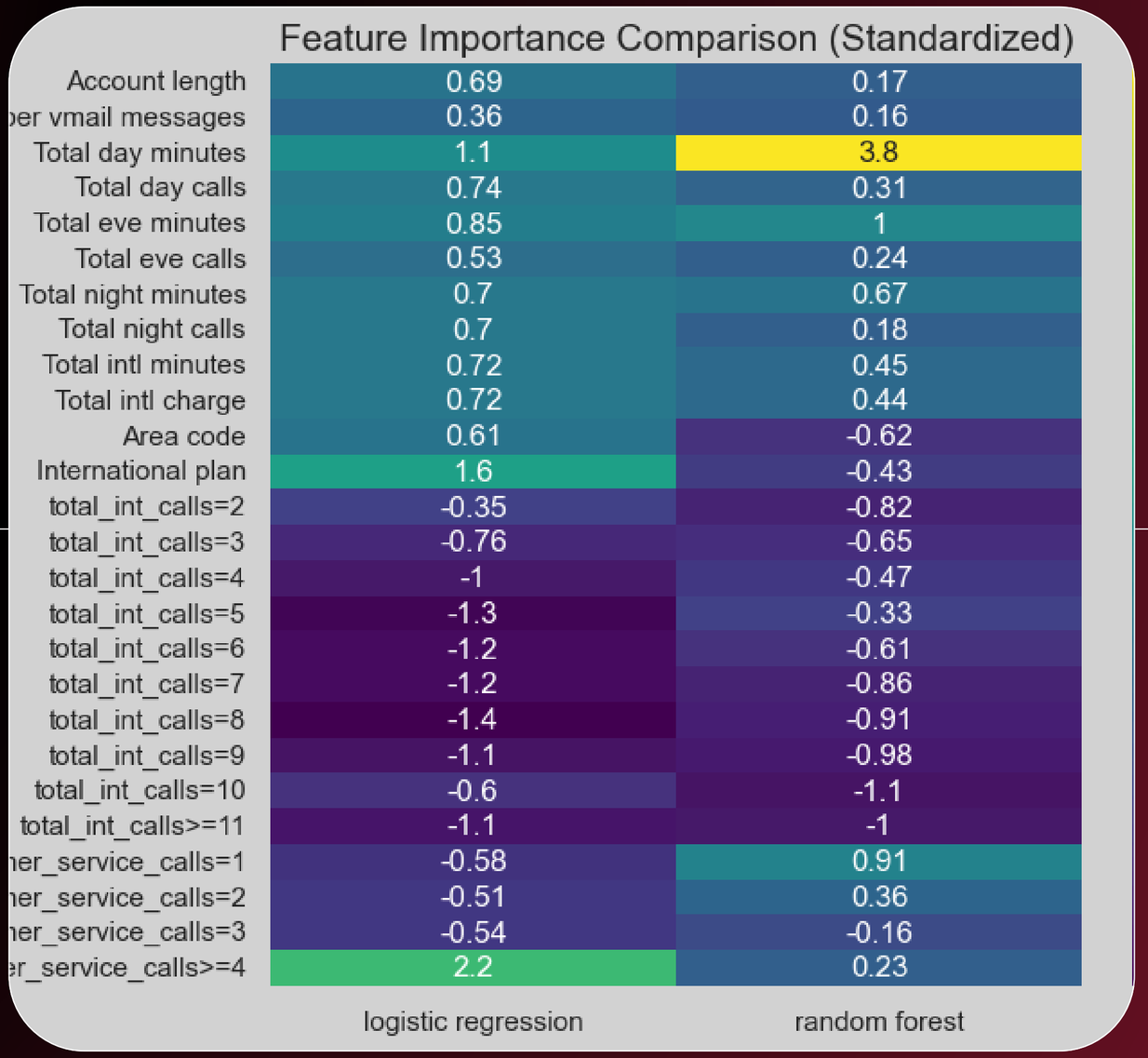


Figure 6: Feature importance comparison between logistic regression and random forest models. Importances for each model were standardized according to the mean and standard deviation of features under the same model.

Results and Recommendations

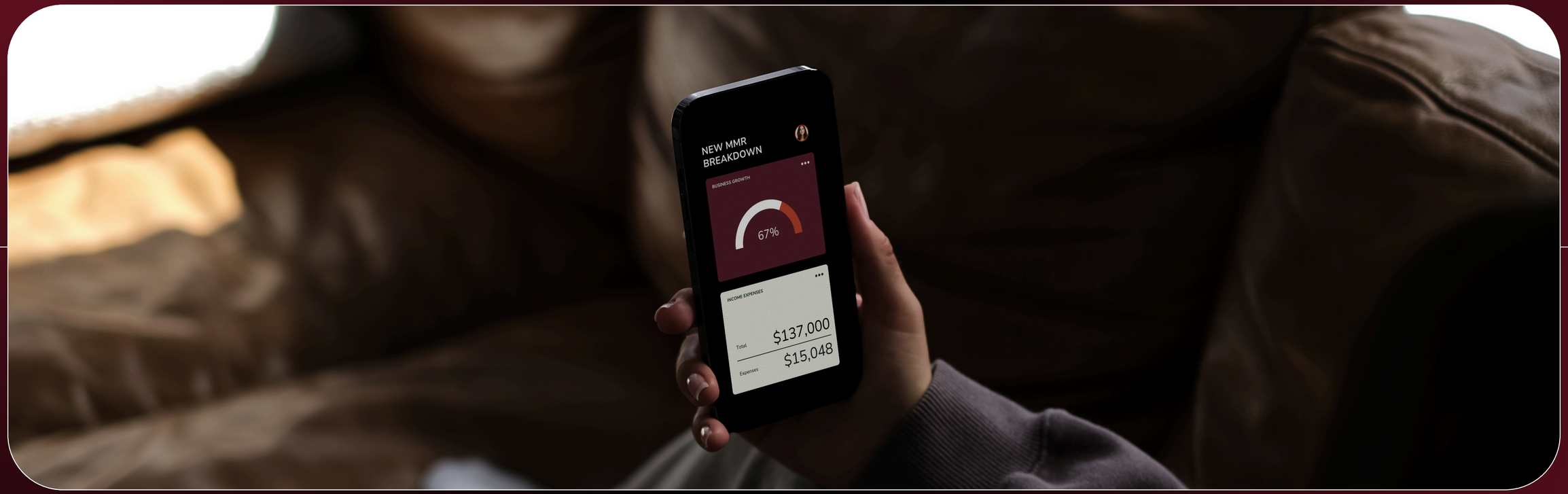
Before making any immense business decision, a discussion of nuances is in order:

PREDICTIVE POWER VS INTERPRETABILITY

It is clear that Random Forest showed superior result in terms of predictive accuracy. However, If the budget is limited and aims to address only few behaviors of customers to make key business decisions, Logistic Regression might just be the better choice since the distinction between feature importance is clearer in Logistic Regression. However, if the company is looking for a specific model to accurately predict or flag a probable churner, the following sections ought to be considered.

BIAS-VARIANCE TRADE-OFF

Random Forest models don't assume relationships between variables, allowing them to detect complex patterns like those in "Total day minutes," while Logistic Regression requires approximately linear relationships, making it simpler but more limited. Random Forest tends toward high variance and overfitting (demonstrated by perfect training accuracy but 18% lower test accuracy), while Logistic Regression leans toward bias and potential underfitting with non-linear data. **For minor performance differences, choose the simpler Logistic Regression; however, with Orange Telecom's Churn Dataset, Random Forest's significantly better performance justifies its selection despite its complexity.**



Conclusion

Although Logistic Regression has the advantage in terms of interpretability (which is beneficial for making key business decisions), a nuanced consideration to variance-bias tradeoff declared the Random Forest as the superior model in terms of predicting churners and non-churners. It thus becomes imperative to ponder upon which aspects of modelling is prioritized, interpretability or maximum predictive performance.

Thankyou

CODE.LINK

PAPER.LINK