Name: Julika Pradhan
Date: 17th May 2024

Customer Churning Prediction and Interpretation

Machine Learning: Churning Prediction

Clustering: Patterns Segmentation to detect risky customers

Objective: Churn Risk Prediction

1. The best churn model is the one that has both predictive and interpretability capabilities.

2. High Recall: Model that can predict the highest True Positives and reduce False Negatives

3. Customer segmentation and extensive data analytics are required to analyze the behavior

Interpretability is about the extent to which a cause and effect can be observed

Predictive is about how well the model can learn and predict from the patterns

Tools and Versions

Framework

1. Jupyter Notebook

Data Visualizations

1. Matplotlib
2. Seaborn

Model Building

1. Sklearn
2. imblearn
3. xgboost

Data Manipulations and Analysis

1. Pandas
2. NumPy
3. SciPy
4. itertools

Versions:

Jupyter Notebook: 4.7.1
Pandas: 1.2.4
NumPy: 1.22.0
SciPy: 1.10.1
Seaborn: 0.11.1
Matplotlib: 3.6.0
Sklearn: 1.2.2
imblearn: 0.12.2
Xgboost: 2.0.1

Data

Sample Size: 7043 rows, 21 columns

Group	Features	Value Sample
	customerID	7590-VHVEG
Demographic	gender	Female / Male
	SeniorCitizen	1 / 0
	Partner	Yes / No
	Dependents	Yes / No
	tenure	1-----------------72
Services	PhoneService	Yes / No
	MultipleLines	Yes / No / No Phone Service
	InternetService	DSL / Fiber Optics / No
	OnlineSecurity	Yes / No / No Internet Service
	OnlineBackup	Yes/No/No Internet Service
	DeviceProtection	Yes / No / No Internet Service
	TechSupport	Yes / No / No Internet Service
	StreamingTV	Yes / No / No Internet Service
	StreamingMovies	Yes / No / No Internet Service
Customer Account	Contract	Month-To-Month / one year / two year
	PaperlessBilling	Yes/No
	PaymentMethod	Electronic check / Mailed check / Bank transfer / Credit Card
	MonthlyCharges	18.25--------118.75
	TotalCharges	18.80--------8684.80
Who has left	Churn	Yes / No

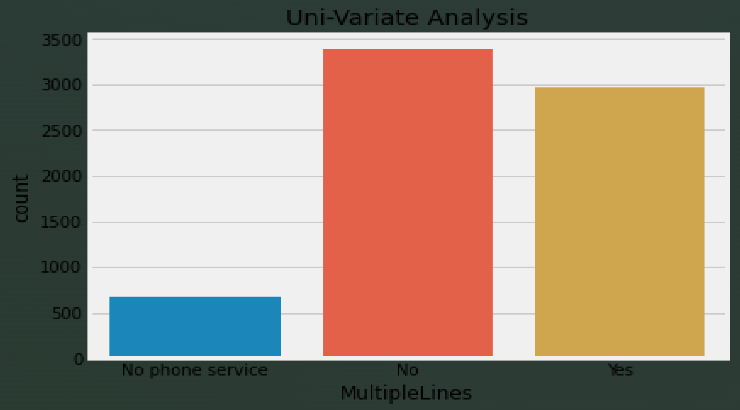
Type of Data: Supervised

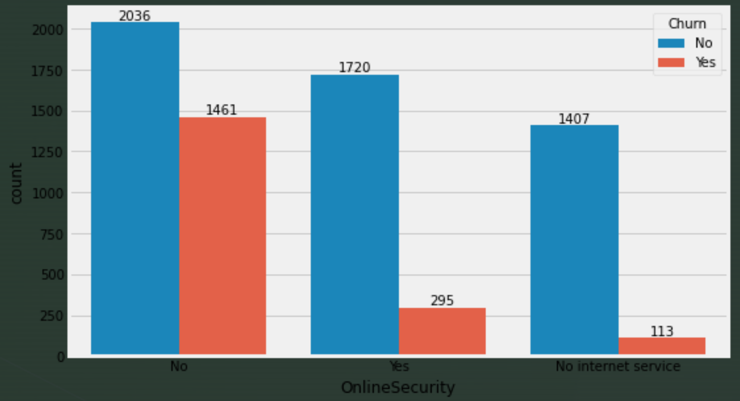
Data Cleaning

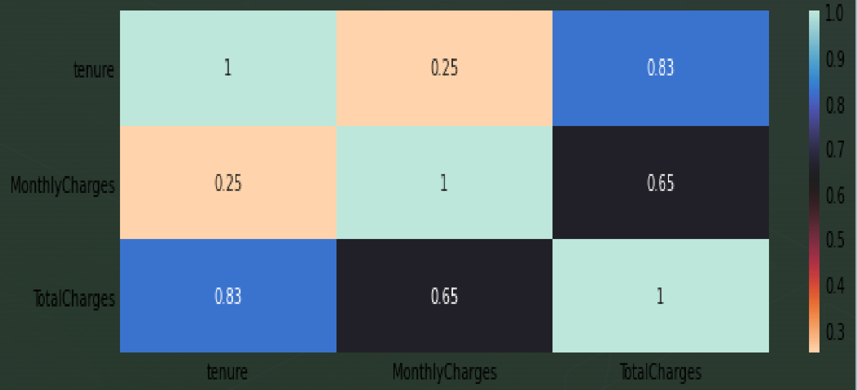
1. Removed Customer ID
2. Converted Data-Type of Total Charges to “float”
3. Removed missing value (0 tenure, 1.5 %)
4. Converted Senior Citizens to Categorical
5. No duplicate data. 22 data seems like duplicates after removing Customer-ID, but they are different customers
6. No presence of outlier

Methods used in EDA

1. Data Manipulation to check certain distributions
2. Churn Distribution Check
3. Uni-variate Analysis
4. Bi-Variate Analysis
5. Multi-Variate Analysis
6. Descriptive Statistics
7. Hypothesis Testing

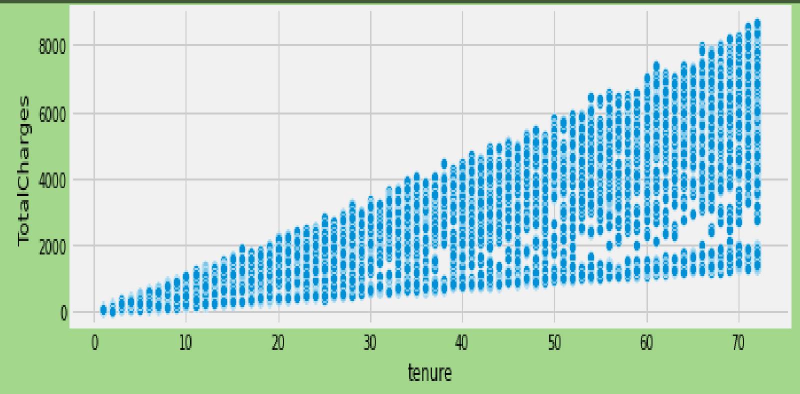


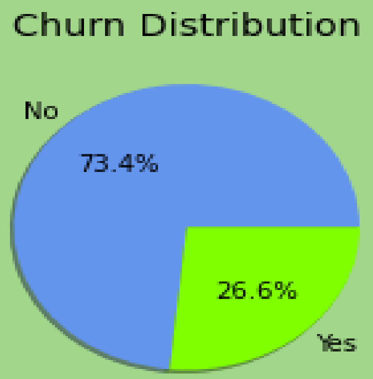


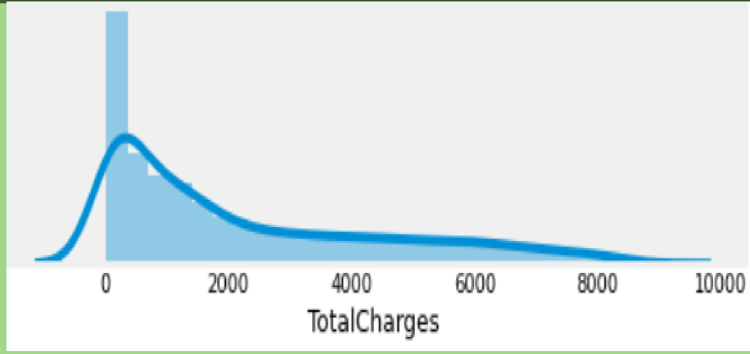


Useful Insights

- Presence of Multi-collinearity.
- Data has both linear and non-linear nature. More features have collinearity with the target.
- Presence of skewness (Example: Total Charges).
- Most features are categorical, so numerical representation is needed.
- Class Imbalance is present.

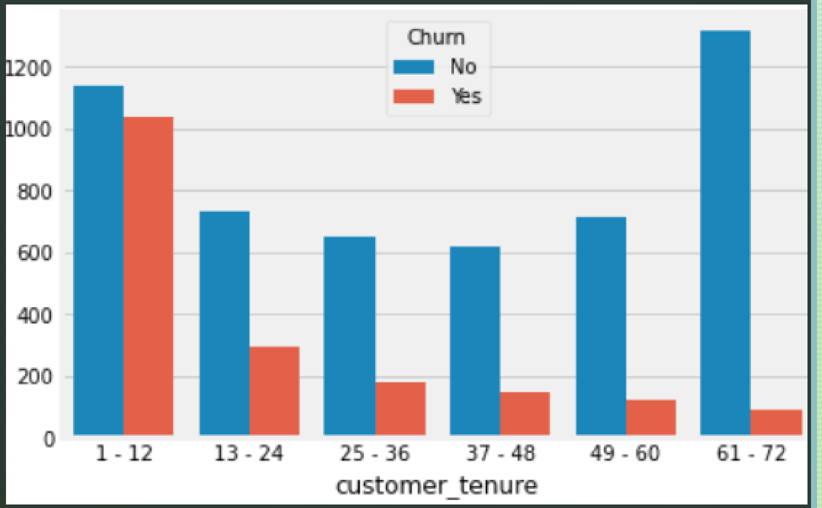






Feature Engineering

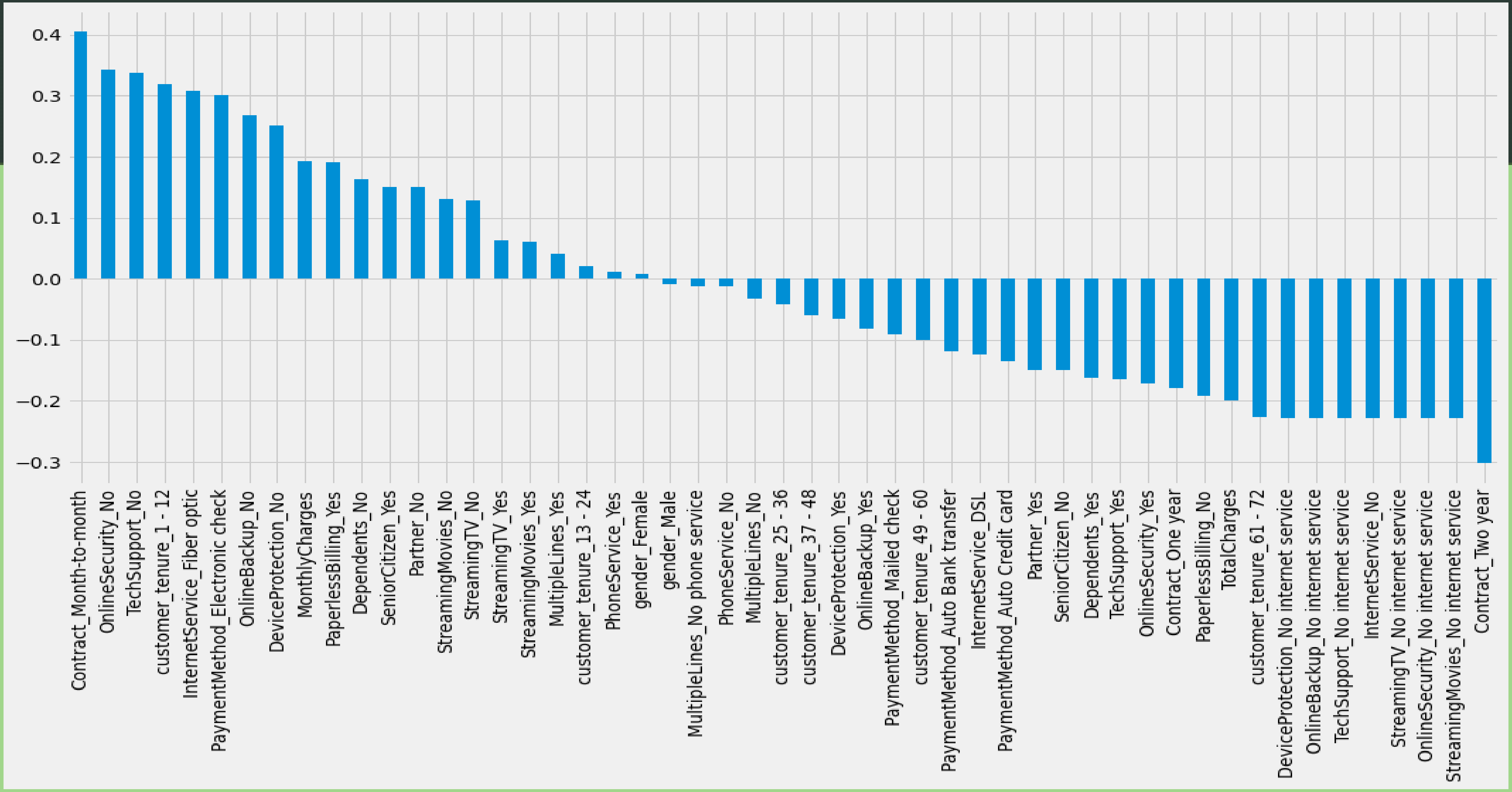
- Binning (Pandas CUT method)
- Encoding (Binary / Sklearn: One-Hot)
- Data Splitting (Sklearn: 25% test, Stratified)
- Sampling (SMOTE)
- Scaling (Standard-Scaler)



MonthlyCharges	TotalCharges
-0.008816	-0.172578
-0.217370	0.674161
1.025677	1.863630
-1.031723	-0.812583
0.903193	1.530876

DeviceProtection_Yes	TechSupport_No	TechSupport_No internet service	TechSupport_Yes	StreamingTV_No	StreamingTV_No internet service	StreamingTV_Yes	StreamingMovies_No
0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0
1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0
1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0

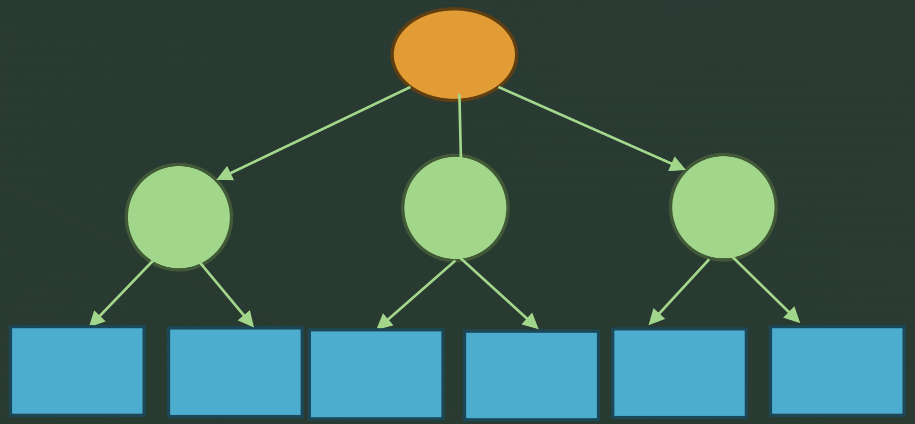
Useful Insights: Correlation with Churn



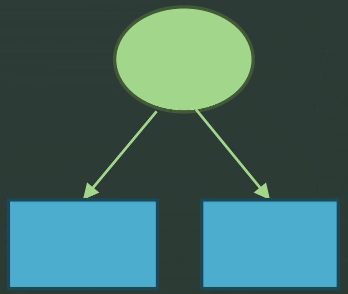
Models and Assumptions

- Linear and Non-Linear
- Can handle Multi-Collinearity
- Feature Importance Capability
- Simple Functionality
- Properties to handle imbalances
- Supervised Classification

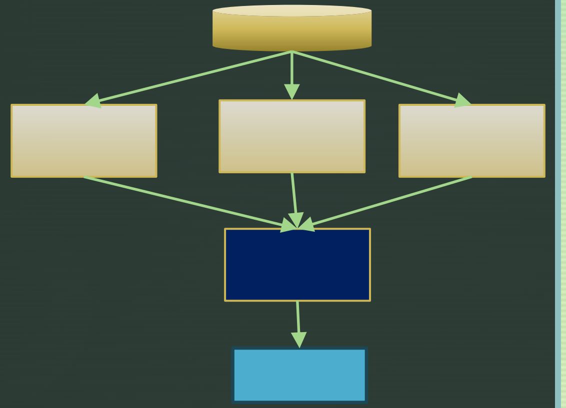
Random Forest



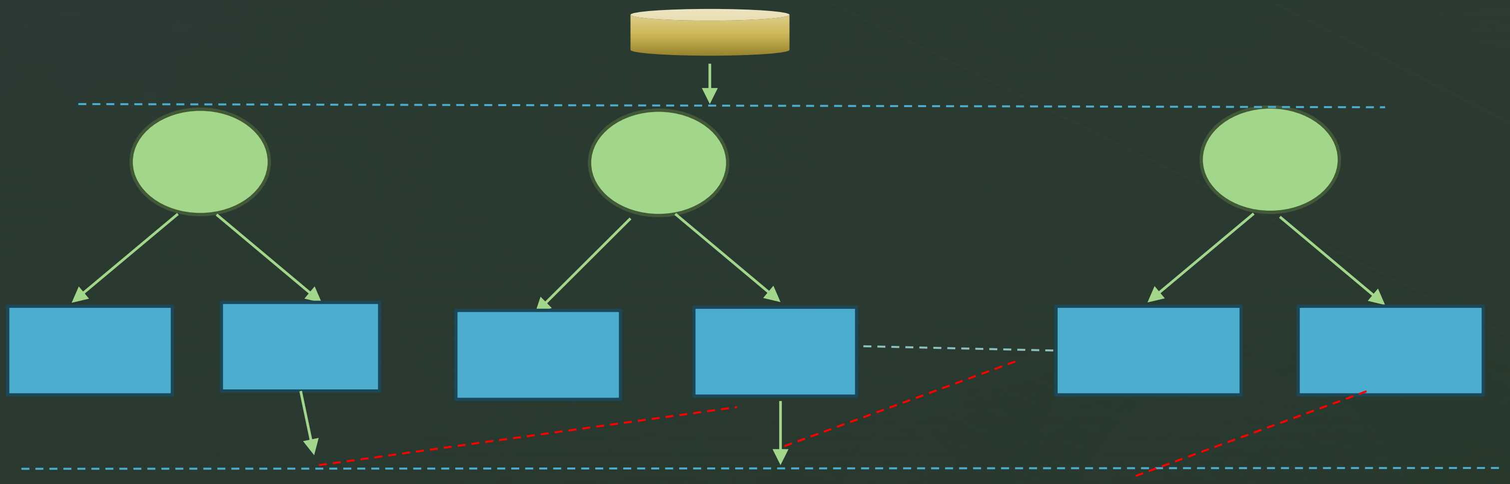
Decision Tree



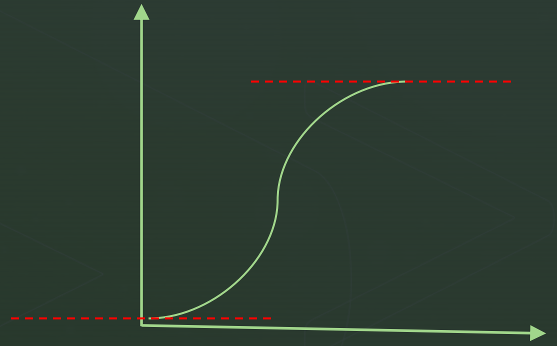
Stacking Model



XGBoost



Logistic Regression



Techniques for Model Building

Decision Tree

Simple Model

DecisionTreeClassifier()

Hyper Parameter Tuning

Method: GridSearchCV

Cross Validation fold: 5
Scoring: Recall

Best Parameters identified:
criterion: gini, max_depth: 10, max_features: log2, min_samples_leaf: 10, min_samples_split: 4

XGBoost

Simple Model

xgb.XGBClassifier()

Hyper Parameter Tuning

Method: GridSearchCV, RandomizedSearchCV, Recursive Feature Elimination (REF)

Cross Validation fold: 5
n_iter (RandomizedSearchCV)= 15
Scoring: Recall

Best Parameters by Grid CV:
'colsample_bytree': 0.9, 'gamma': 0.0, 'learning_rate': 0.001, 'max_delta_step': 1, 'max_depth': 4, 'min_child_weight': 4, 'n_estimators': 1000, 'nthread': 4, 'objective': 'binary: logistic', 'scale_pos_weight': 2.7, 'subsample': 0.9

Best Parameters by Random CV:
'colsample_bytree': 0.515087, 'gamma': 0.1179641868, 'learning_rate': 0.002203275702028241, 'loss': 'deviance', 'max_delta_step': 1, 'max_depth': 5, 'min_child_weight': 8, 'n_estimators': 954, 'nthread': 4, 'objective': 'binary: logistic', 'scale_pos_weight': 2.7, 'subsample': 0.8475254937

BalancedRandomForest

Simple Model

BalancedRandomForest()

Hyper Parameter Tuning

Method: RandomizedSearchCV, Recursive Feature Elimination CV (RFECV)

Cross Validation fold: 5
RepeatedStraifiedKfold: splits-5, repeat-3 (For RFECV)
n_iter (RandomizedSearchCV)= 15
Scoring: Recall

Best Parameters by Random CV :
'n_estimators': 1000, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'log2', 'max_depth': 10

Best Parameters by using RFECV features with Random CV:
'n_estimators': 300, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 10

Logistic Regression

Simple Model

LogisticRegression()

Hyper Parameter Tuning

Method: GridSearchCV

RepeatedStraifiedKfold: splits-10, repeat-7
Scoring: Recall

Best Parameters by Grid CV:
'C': 0.001, 'class_weight': 'balanced', 'max_iter': 10, 'penalty': 'l2', 'solver': 'liblinear'

Model Stacking

Base Models: Random Forest, Decision Tree, and XGBoost models

Meta Model: Above specified tuned Logistic Regression

Model Performance

	Pred_Accuracy	Pred_Recall	Pred_precision	Pred_F1
DecisionTree	0.74	0.51	0.50	0.50
DecisionTree-GridCV	0.78	0.62	0.43	0.51
DecisionTree-SMOTE	0.77	0.55	0.60	0.58
XGBoost	0.77	0.58	0.49	0.54
XGBoost-GridCV	0.73	0.82	0.49	0.62
XGBoost-GridCV-Imp-Features	0.73	0.82	0.49	0.62
XGBoost-RandomCV	0.74	0.81	0.50	0.62
BalancedRandomForest	0.72	0.79	0.49	0.60
BalancedRandomForest-GridCV	0.72	0.80	0.49	0.61
BalancedRandomForest-GridCV_Imp-Features	0.73	0.81	0.50	0.62
LogisticRegression	0.79	0.50	0.65	0.56
LogisticRegression-SMOTE	0.74	0.79	0.51	0.62
LogisticRegression-StandardScaler	0.72	0.80	0.48	0.60
Model-Stacking	0.70	0.84	0.46	0.60

Goal: Identify truly predicted churned customers and reduce False Negative

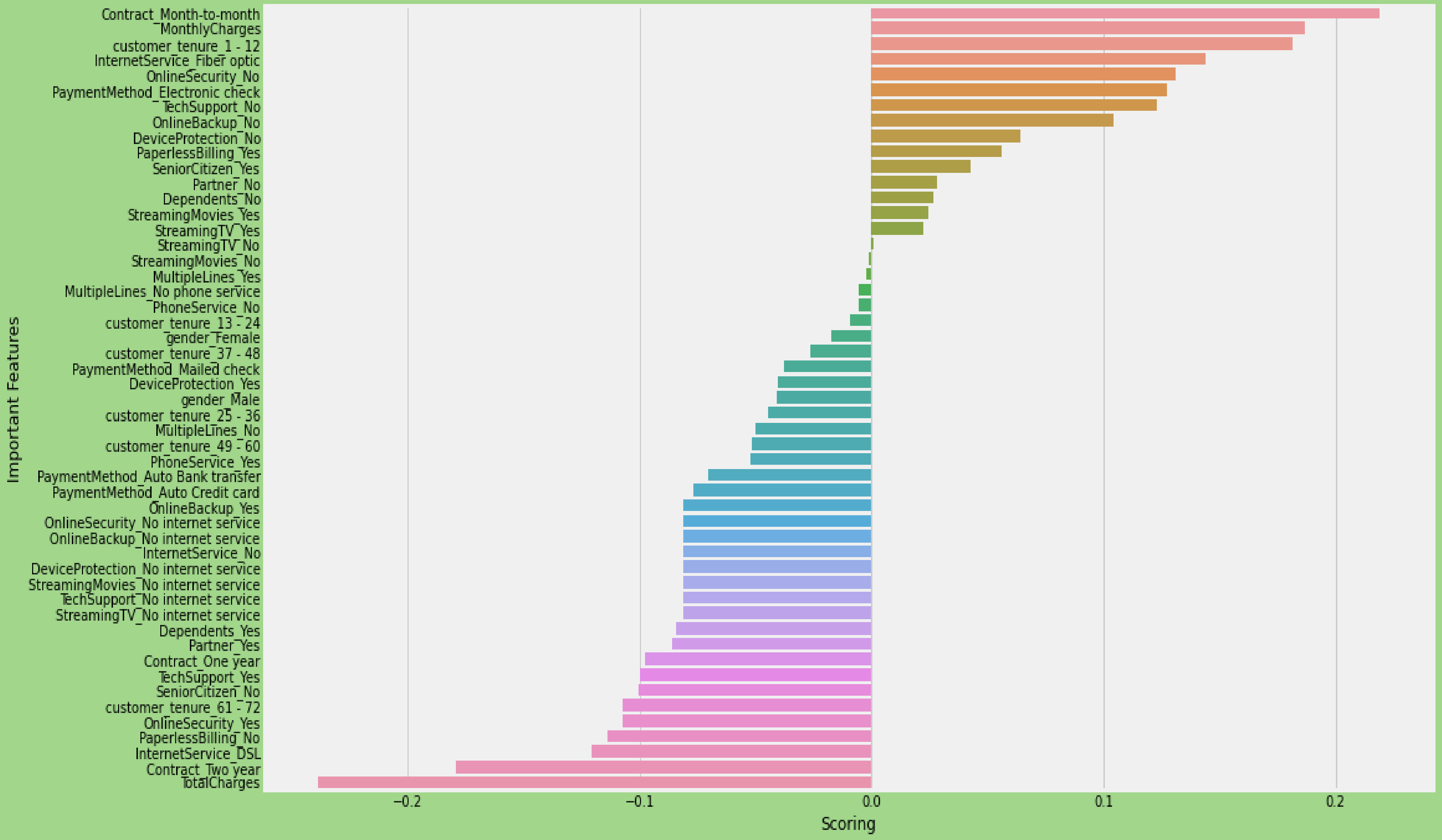
Important Metric to Consider: Recall

Models with >= 80% recall: Tuned XGBoost, Tuned Random Forest, Tuned Logistic Regression, Stacked Models

Best Model-84%: Model Stacking (Best Predictor: XGBoost)

Feature Importance; Stack Models

Feature Importance: Logistic Regression



Model Summary

- Stacking Model has the highest churning prediction of, 84%.

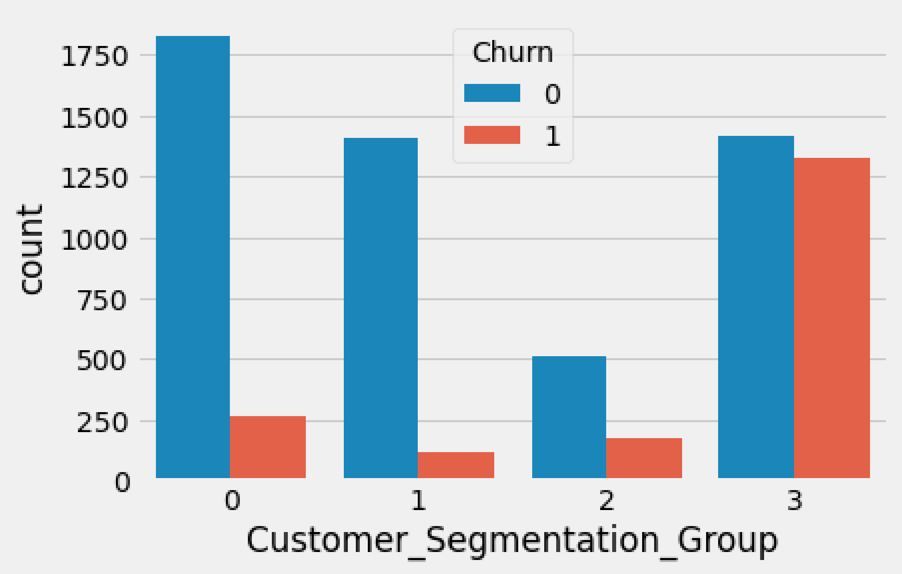
- Logistic Regression alone has an 80% Recall rate.

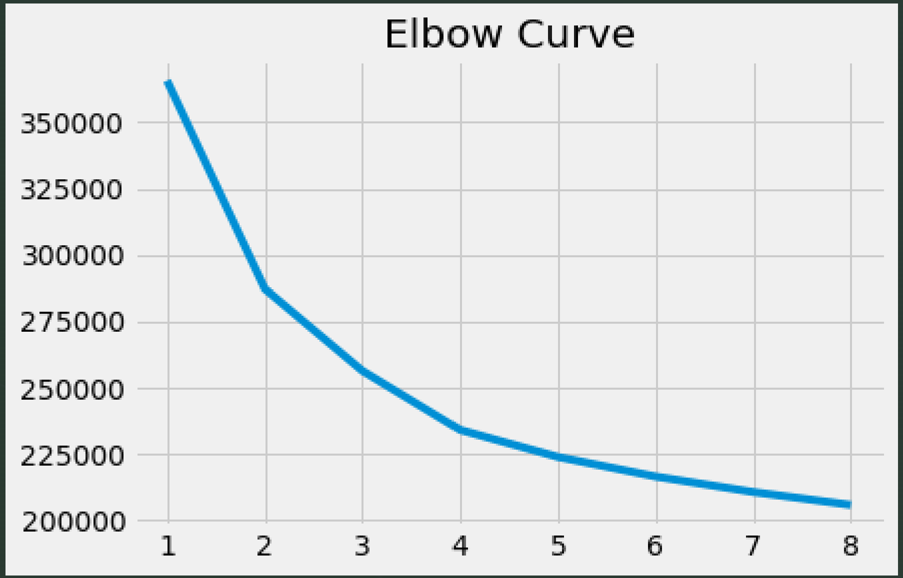
- However, the feature importance, which is crucial for a churn project that describes the model's interpretability by identifying risk factors, is not presented. If we compare the correlations and segmentation analysis, the feature importances extracted are not completely aligned with what these models have interpreted.

- The feature importances extracted by the Logistic Regression model seem to be very clear indicating the model's high interpretability.

Customer Segmentation

Model: K-Means





Segmentation group – 3 has at least 50% of the churned customers. So, company should pay attention to the non-churned customers present in this group because they share similar behaviour. This segment of customers is more likely to be at risk.

Summary

Churning Drivers by Logistic Regression

1. Contract: Month-to-month
2. Tenure: Newest customers
3. Internet service: Fiber optic
4. Payment method: Electronic check
5. Charged: Monthly
6. Having No access to services: online security, tech support, device protection, online backup
7. No partners
8. No dependents
9. Paperless billing

Future Work

More hyperparameter experimentation on Logistic Regression to improve performance

More research on the Stacking Model to improve its interpretability