Project Group B5
Clarissa Lee Zixin
Dong Mingjie
Fan Zixian
Song Yida

Telco Customer Churn Prediction

Introduction

Motivation,
Problem Description,
Goal

Data

Overview,
pre-processing
feature selection

Models

Model training ,
Accuracy,
comparison

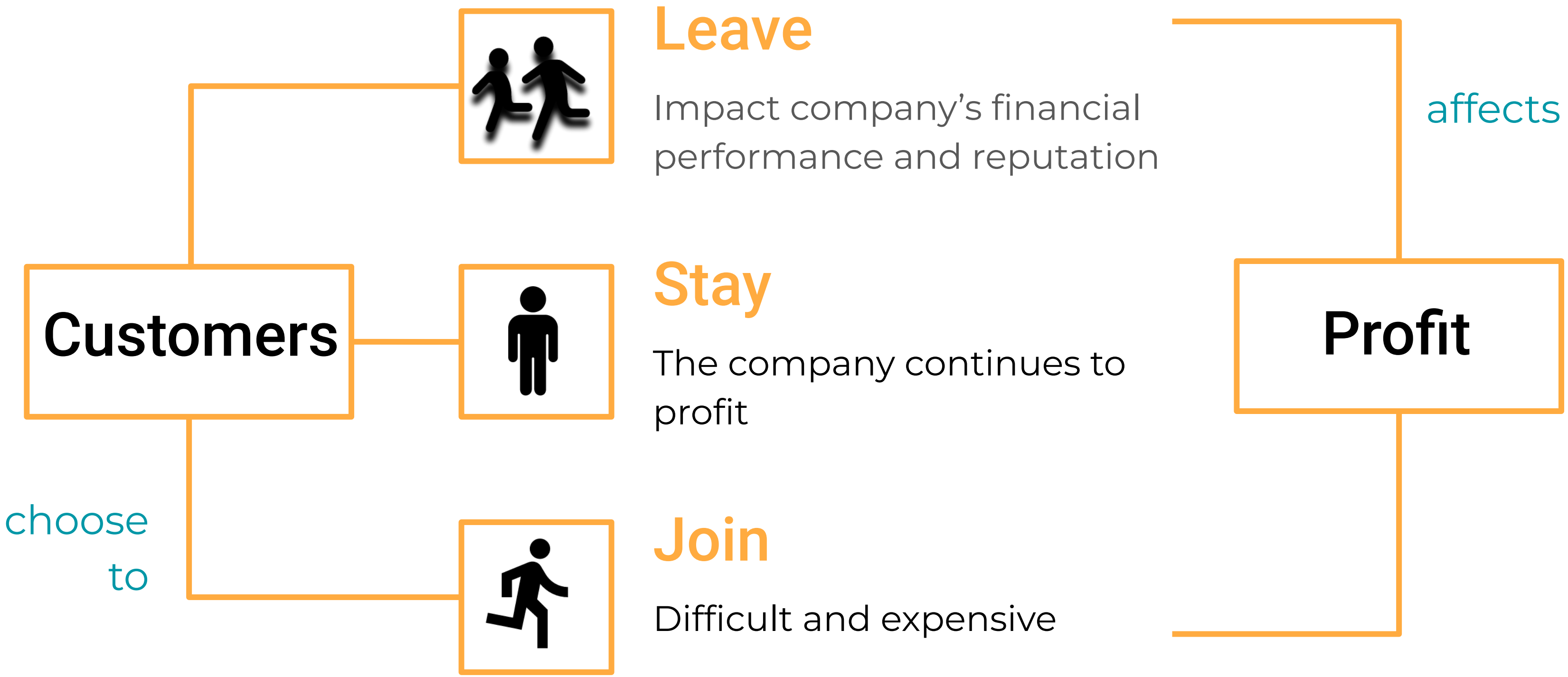
Conclusion

Summary,
Recommendation

1. Introduction

Motivation, Problem Description, Goal

Motivation


Leave
Impact company’s financial performance and reputation
Stay
The company continues to profit
Join
Difficult and expensive
Customers
choose to
Profit
affects

Problem statement

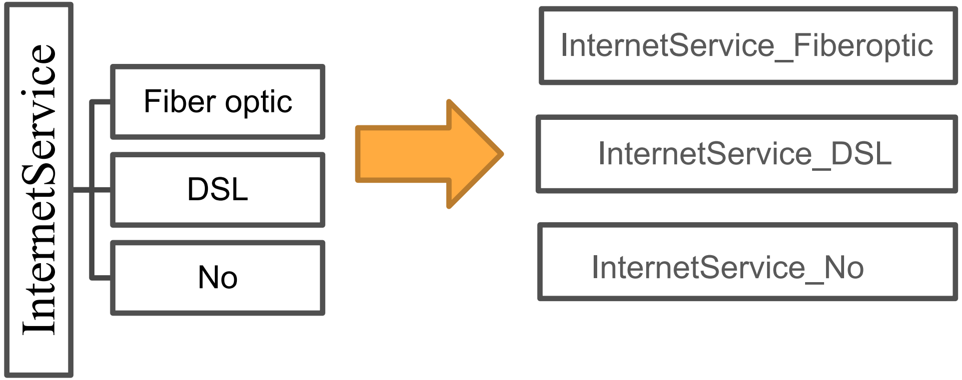
Using statistical learning methods to build
a predictive model
which helps to identify customers that are
most likely to churn based on the
customers’ information.

2. Data

Overview; Pre-processing; Feature selection

Main features

Split some variables



7032 × **26**

7032 entries, each represent a customer

26 features, containing customer attributes, services, account information, and demographics

Multicollinearity

	feature	VIF
0	gender	1.002106
1	SeniorCitizen	1.153220
2	Partner	1.462988
3	Dependents	1.381598
4	tenure	7.584453
5	PhoneService	34.893857
6	MultipleLines	7.289761
7	OnlineSecurity	6.338349
8	OnlineBackup	6.796678
9	DeviceProtection	6.924754
10	TechSupport	6.476508
11	StreamingTV	24.080019
12	StreamingMovies	24.156394
13	PaperlessBilling	1.208455
14	MonthlyCharges	866.089640
15	TotalCharges	10.811490
16	InternetService_DSL	inf
17	InternetService_Fiber optic	inf
18	InternetService_No	inf
19	Contract_Month-to-month	inf
20	Contract_One year	inf
21	Contract_Two year	inf
22	PaymentMethod_Bank transfer (automatic)	inf
23	PaymentMethod_Credit card (automatic)	inf
24	PaymentMethod_Electronic check	inf
25	PaymentMethod_Mailed check	inf

- Some of the variables have $\boxed{\text{infinity}}$ returned (perfectly correlated)
- PhoneService, StreamingTV, StreamingMovies, MonthlyCharges and TotalCharges have high VIF value of more than 10

Data Scaling

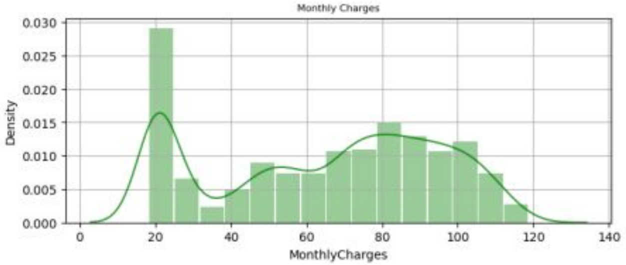


Figure 4: MonthlyCharges Histogram Display

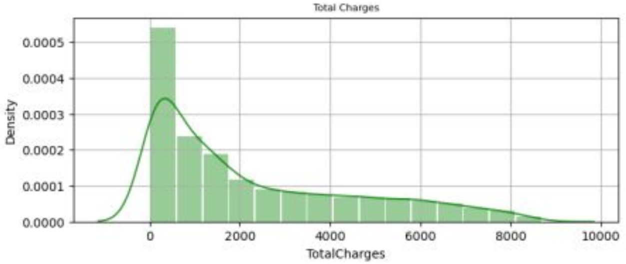
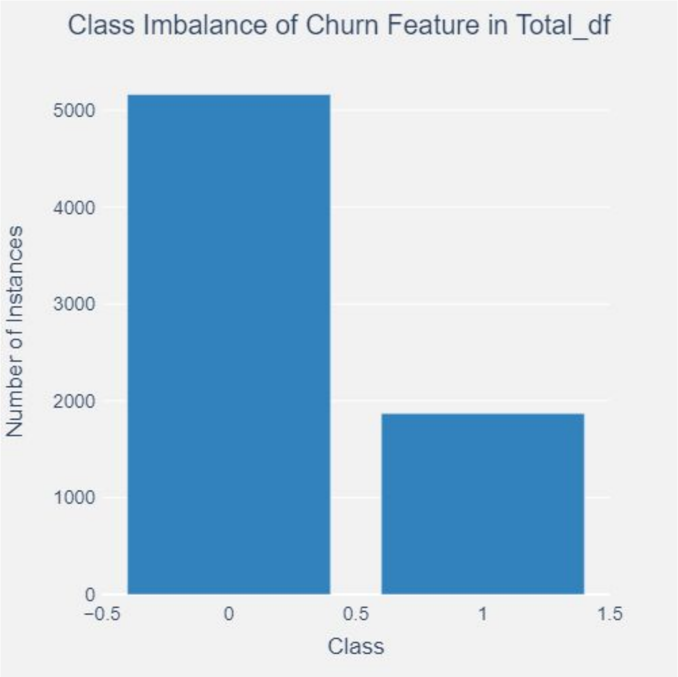


Figure 5: TotalCharges Histogram Display

- There are some non-normal behaviour in some quantitative variables

- We consider standard scaler, min-max scaler and quantile scaler on different models

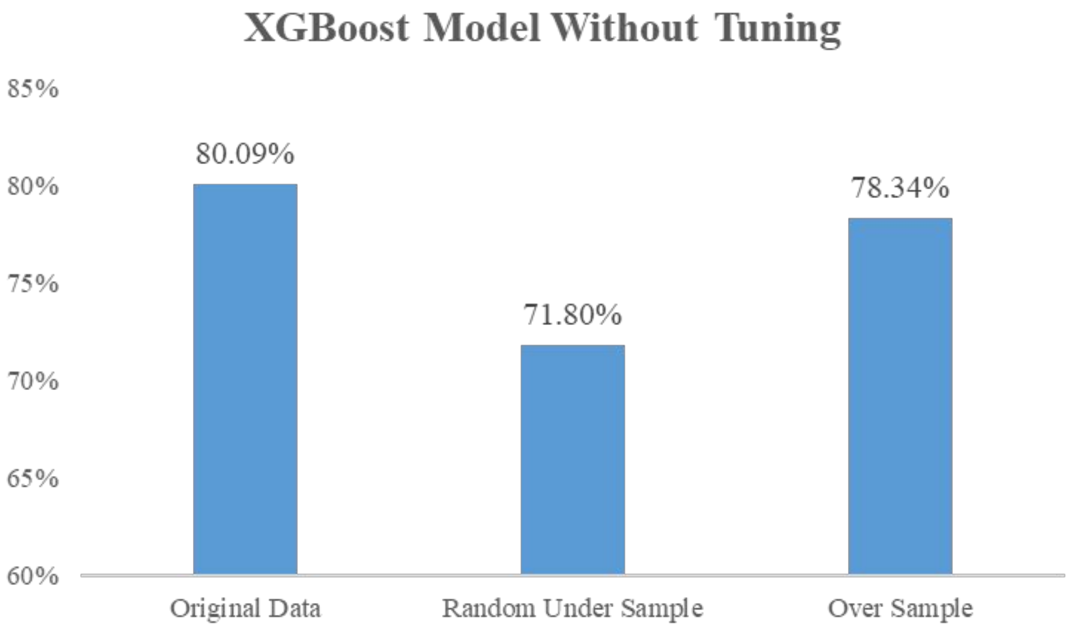
Imbalanced Dataset



- The model fitted by general methods will skew to the majority class "No"

- For the explanatory variable "Churn": "No" label is about 3 times larger than "Yes" labels

Imbalanced Dataset

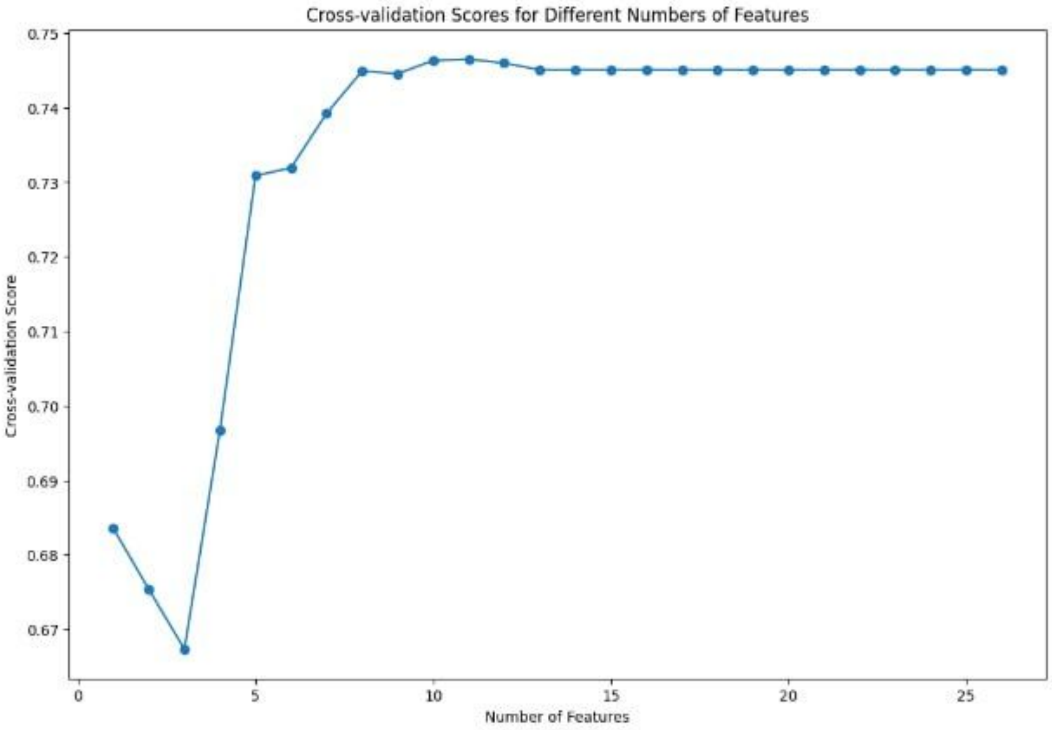


- Random-under-sample and over-sample are used to handle this imbalance
- However, both treatments performed poorly

Reasons for poor performance of balanced methods

- According to He, H., & Garcia, E. A. (2009), the imbalance ratio below 1:10 is considered to be insignificant
- Hence, the 3:1 ratio of “No” to “Yes” in our dataset is not a severe imbalance problem
- The imbalanced distribution is reflective of that in the real world at 3:1
- Losing information

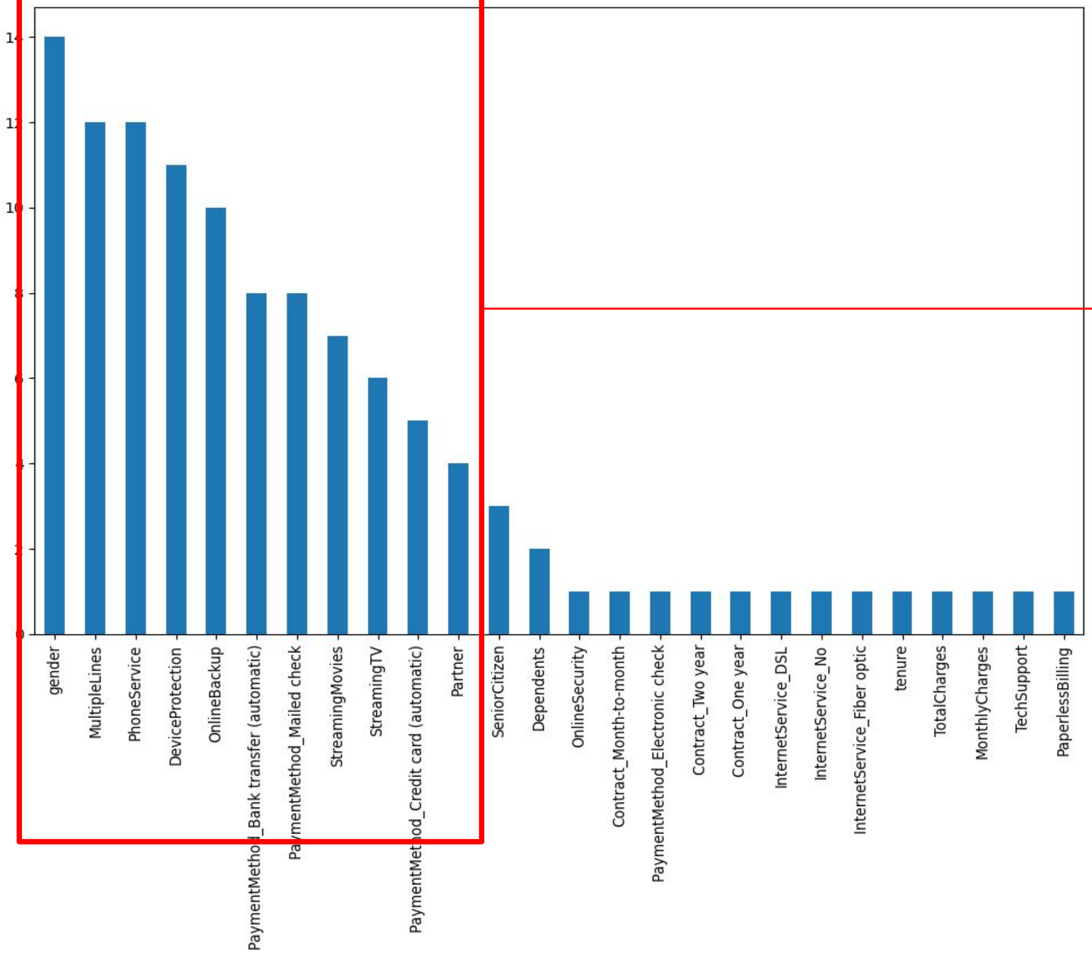
Feature selection




```
The best number of features is: 11
```

- We select important features by Boruta
- 11 in 26 features are selected via cross-validation

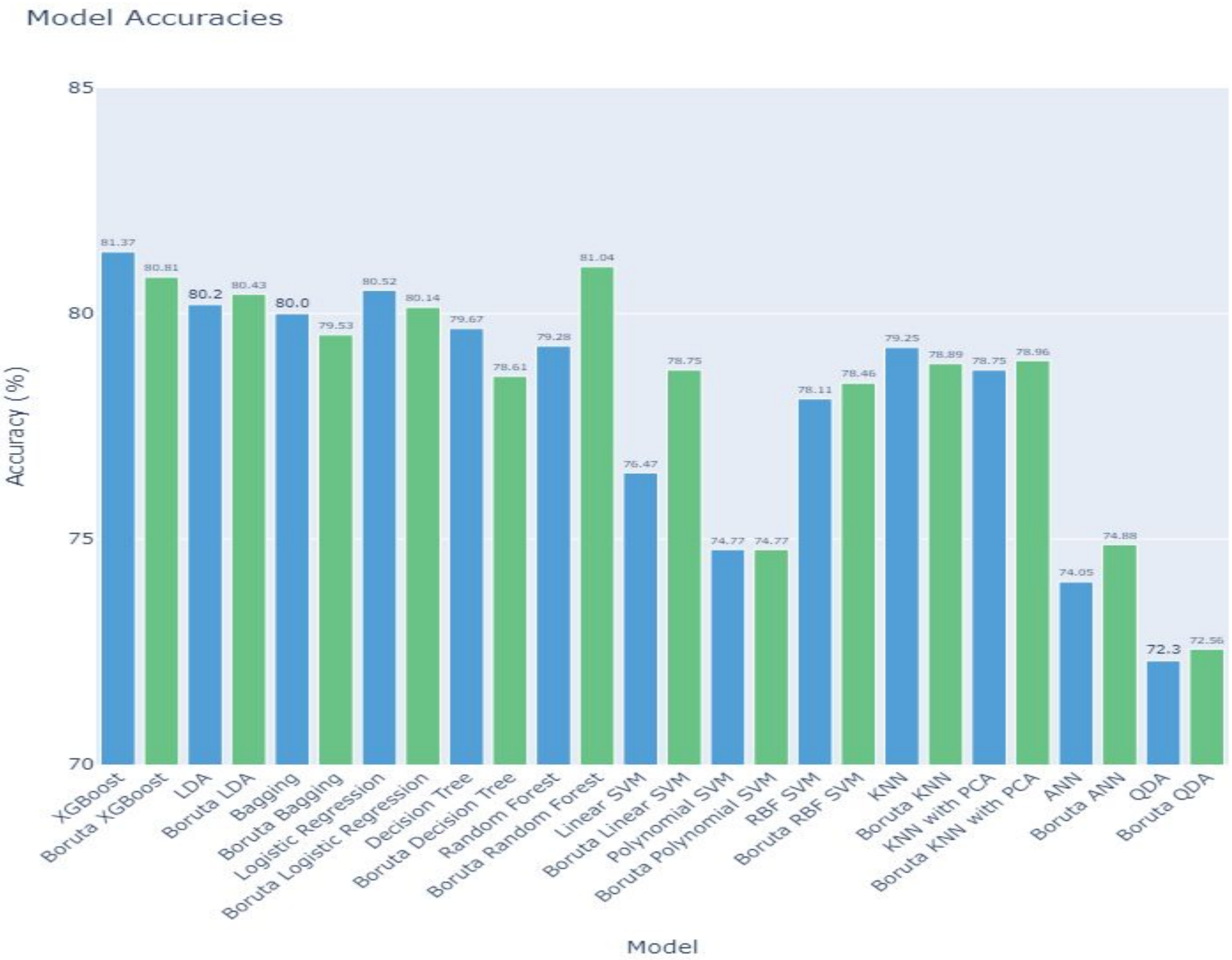
Feature selection



'tenure', 'OnlineSecurity', 'TechSupport', 'PaperlessBilling', 'MonthlyCharges', 'TotalCharges', 'InternetService_DSL', 'InternetService_Fiber optic', 'InternetService_No', 'Contract_Month-to-month', 'Contract_One year'.

3. Model

Model training; Accuracy; Comparison


Model Accuracies
Accuracy (%)
85
80
75
70
81.37
80.81
80.2
80.43
80.0
79.53
80.52
80.14
79.67
78.61
79.28
81.04
76.47
78.75
74.77
74.77
78.11
78.46
79.25
78.89
78.75
78.96
74.05
74.88
72.3
72.56
XGBoost
Boruta XGBoost
LDA
Boruta LDA
Bagging
Boruta Bagging
Logistic Regression
Boruta Logistic Regression
Decision Tree
Boruta Decision Tree
Random Forest
Boruta Random Forest
Linear SVM
Boruta Linear SVM
Polynomial SVM
Boruta Polynomial SVM
RBF SVM
Boruta RBF SVM
KNN
Boruta KNN
KNN with PCA
Boruta KNN with PCA
ANN
Boruta ANN
QDA
Boruta QDA
Model

3.1

Logistic Regression

Logistic Regression

Fit with Initial scaled dataset

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-0.1225	0.1680	-0.7287	0.4662	-0.4518	0.2069
gender	0.0225	0.0776	0.2900	0.7718	-0.1296	0.1746
SeniorCitizen	0.2769	0.1021	2.7126	0.0067	0.0768	0.4770
Partner	-0.0212	0.0935	-0.2271	0.8203	-0.2046	0.1621
Dependents	-0.1135	0.1082	-1.0485	0.2944	-0.3255	0.0986
PhoneService	0.7239	0.7829	0.9246	0.3552	-0.8106	2.2584
MultipleLines	0.5206	0.2146	2.4262	0.0153	0.1000	0.9411
OnlineSecurity	-0.0753	0.2157	-0.3491	0.7270	-0.4981	0.3475
OnlineBackup	0.1988	0.2127	0.9349	0.3499	-0.2180	0.6157
DeviceProtection	0.2476	0.2115	1.1708	0.2417	-0.1669	0.6620
TechSupport	-0.0860	0.2158	-0.3984	0.6903	-0.5089	0.3369
StreamingTV	0.9323	0.3958	2.3556	0.0185	0.1566	1.7080
StreamingMovies	0.7770	0.3943	1.9706	0.0488	0.0042	1.5498
PaperlessBilling	0.2710	0.0889	3.0489	0.0023	0.0968	0.4452
InternetService_DSL	-0.0547	0.0878	-0.6235	0.5329	-0.2268	0.1173
InternetService_Fiber_optic	2.3949	1.0074	2.3774	0.0174	0.4205	4.3693
InternetService_No	-2.4626	0.9233	-2.6672	0.0076	-4.2723	-0.6530
Contract_Month	0.7104	0.0949	7.4867	0.0000	0.5245	0.8964
Contract_One_year	-0.0490	0.1107	-0.4425	0.6581	-0.2661	0.1680
Contract_Two_year	-0.7839	0.1590	-4.9293	0.0000	-1.0956	-0.4722
PaymentMethod_Bank_transfer	-0.0707	0.0924	-0.7654	0.4440	-0.2518	0.1104
PaymentMethod_Credit_card	-0.1400	0.0965	-1.4513	0.1467	-0.3291	0.0491
PaymentMethod_Electronic_check	0.2203	0.0743	2.9633	0.0030	0.0746	0.3660
PaymentMethod_Mailed_check	-0.1320	0.0907	-1.4562	0.1453	-0.3098	0.0457
tenure	-3.9702	0.5276	-7.5253	0.0000	-5.0042	-2.9362
TotalCharges	2.3961	0.7328	3.2696	0.0011	0.9598	3.8324
MonthlyCharges	-6.4828	3.8442	-1.6864	0.0917	-14.0174	1.0517

- Test accuracy is 80.85%
- Insignificant variables with high p-value (i.e. gender, Partner, etc.)
- Some variables with high p-value were previously highlighted to have high VIF values

	feature	VIF
0	gender	1.002106
1	SeniorCitizen	1.153220
2	Partner	1.462988
3	Dependents	1.381598
4	tenure	7.584453
5	PhoneService	34.893857
6	MultipleLines	7.289761
7	OnlineSecurity	6.338349
8	OnlineBackup	6.796678
9	DeviceProtection	6.924754
10	TechSupport	6.476508
11	StreamingTV	24.080019
12	StreamingMovies	24.156394
13	PaperlessBilling	1.208455
14	MonthlyCharges	866.089640
15	TotalCharges	10.811490

VIF

Logistic Regression

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-1.5955	0.1151	-13.8669	0.0000	-1.8210	-1.3700
SeniorCitizen	0.2936	0.0998	2.9412	0.0033	0.0980	0.4893
MultipleLines	0.3313	0.1050	3.1543	0.0016	0.1254	0.5371
OnlineSecurity	-0.2657	0.1068	-2.4881	0.0128	-0.4750	-0.0564
TechSupport	-0.2668	0.1090	-2.4473	0.0144	-0.4804	-0.0531
StreamingTV	0.5642	0.1167	4.8363	0.0000	0.3356	0.7929
StreamingMovies	0.4156	0.1150	3.6154	0.0003	0.1903	0.6409
PaperlessBilling	0.2773	0.0887	3.1257	0.0018	0.1034	0.4512
InternetService_DSL	-0.5359	0.0891	-6.0118	0.0000	-0.7106	-0.3612
InternetService_Fiber_optic	0.9812	0.1746	5.6202	0.0000	0.6390	1.3234
InternetService_No	-2.0408	0.2159	-9.4544	0.0000	-2.4638	-1.6177
Contract_Month	0.2214	0.0901	2.4579	0.0140	0.0449	0.3980
Contract_One_year	-0.5403	0.0998	-5.4155	0.0000	-0.7358	-0.3447
Contract_Two_year	-1.2766	0.1527	-8.3591	0.0000	-1.5760	-0.9773
PaymentMethod_Bank_transfer	-0.2924	0.1115	-2.6225	0.0087	-0.5109	-0.0739
PaymentMethod_Credit_card	-0.3582	0.1176	-3.0466	0.0023	-0.5886	-0.1277
PaymentMethod_Mailed_check	-0.3570	0.1168	-3.0568	0.0022	-0.5859	-0.1281
tenure	-1.3875	0.1808	-7.6749	0.0000	-1.7419	-1.0332
TotalCharges	0.6513	0.1912	3.4065	0.0007	0.2766	1.0260
MonthlyCharges	-0.8344	0.2060	-4.0512	0.0001	-1.2380	-0.4307

Remove insignificant variables and high VIF variables

- Test accuracy is 80.52%
- Predictors all have small p-value near to 0

Logistic Regression

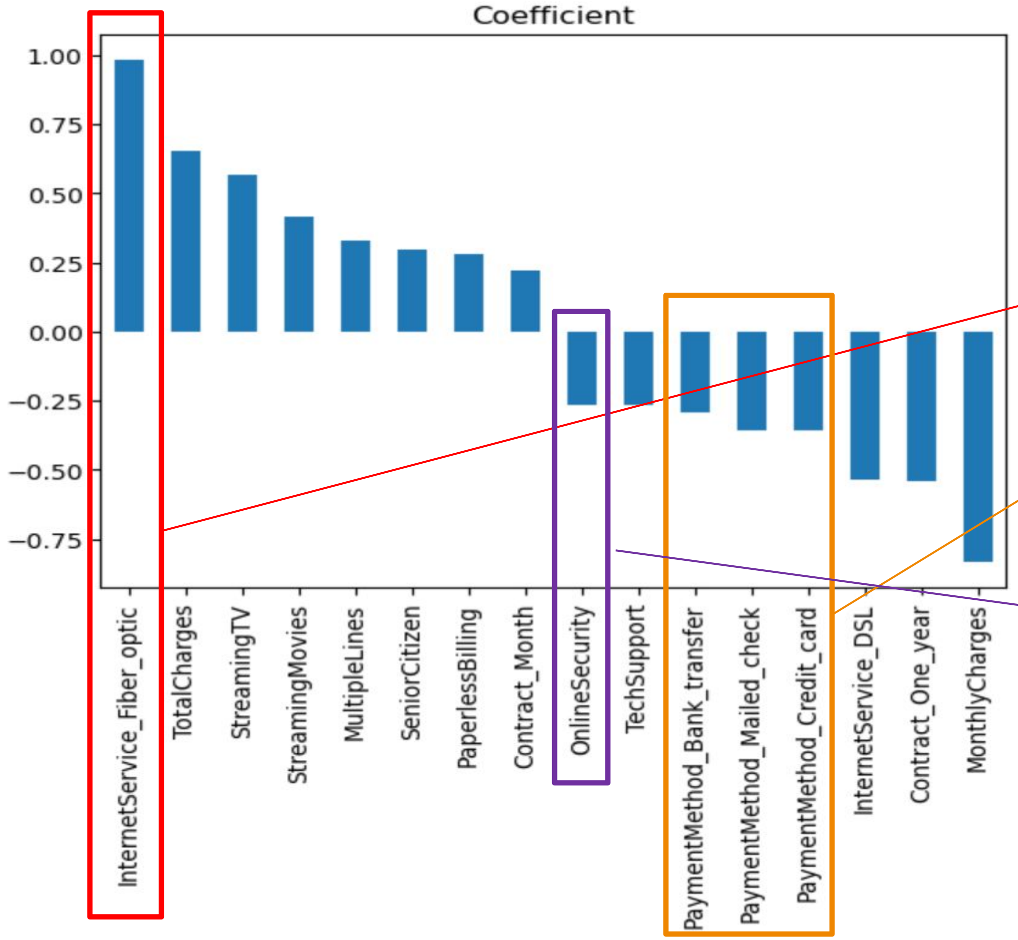
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-2.1643	0.1614	-13.4104	0.0000	-2.4806	-1.8480
tenure	-1.3657	0.1804	-7.5717	0.0000	-1.7192	-1.0122
OnlineSecurity	-0.5031	0.1007	-4.9952	0.0000	-0.7005	-0.3057
TechSupport	-0.4823	0.1042	-4.6303	0.0000	-0.6864	-0.2781
PaperlessBilling	0.3554	0.0872	4.0779	0.0000	0.1846	0.5263
MonthlyCharges	0.1362	0.1315	1.0353	0.3005	-0.1216	0.3940
TotalCharges	0.6779	0.1913	3.5436	0.0004	0.3029	1.0528
InternetService_DSL	-0.6005	0.0889	-6.7554	0.0000	-0.7747	-0.4262
InternetService_Fiber_optic	0.0846	0.1454	0.5818	0.5607	-0.2004	0.3695
InternetService_No	-1.6484	0.1682	-9.8028	0.0000	-1.9780	-1.3188
Contract_Month	1.6036	0.2141	7.4880	0.0000	1.1838	2.0233
Contract_One_year	0.7666	0.2177	3.5219	0.0004	0.3400	1.1931

Fit with selected 11 features in previous feature selected stage

- Test accuracy is 80.14%
  Perform Worse

- Still some Predictors have high p-values(MonthlyCharges and InternetService_Fiber_optic)

Logistic Regression



Model Interpretation

- We decide that the 2nd model performs the best
- InternetService_Fiber_optic has the greatest positive coefficient
- Payment methods matter less to customers churn
- OnlineSecurity has the smallest coefficient

3.2

Support Vector Machine (SVM)

We are using three types of SVM

Linear SVM

- Computationally efficient, even for large datasets
- Performs well in high-dimensional spaces

Accuracy: Original dataset: 76.47%
Boruta dataset: 78.75%

Polynomial SVM

- Can capture complex non-linear relationships between input features and target variable
- Effective when data is not linearly separable and has a complex decision boundary

Accuracy: Original dataset: 74.77%
Boruta dataset: 74.77%

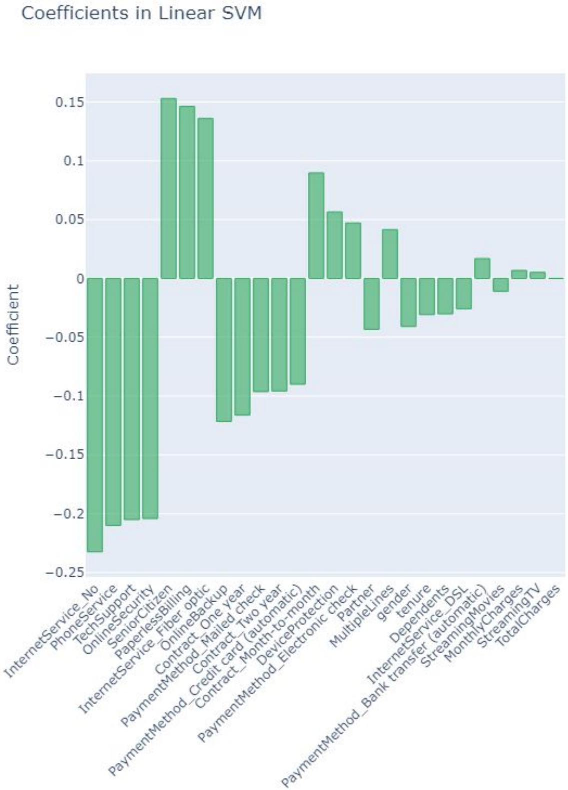
Radial Basis Function (RBF) SVM

- Can capture complex non-linear relationships between input features and target variable
- Effective when data is not linearly separable and has a non-linear decision boundary.

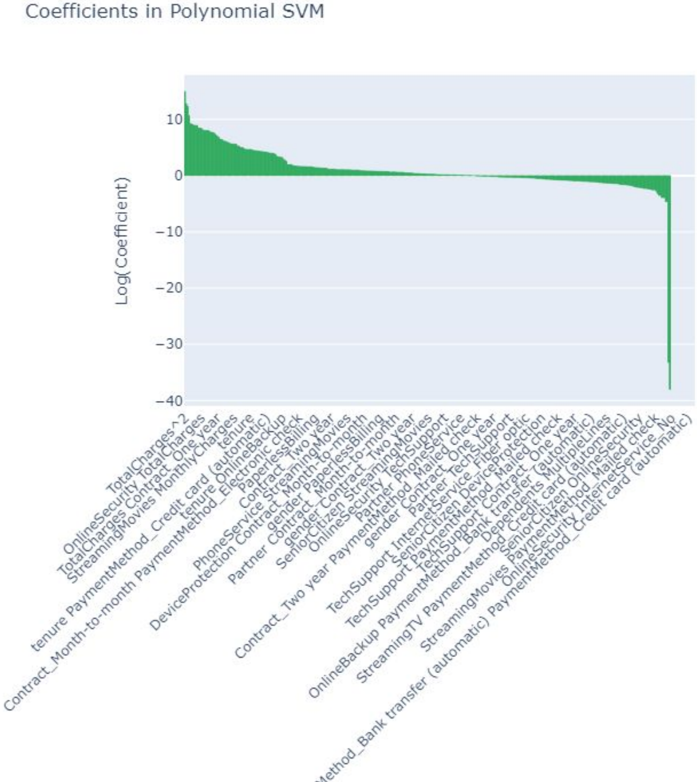
Accuracy: Original dataset: 78.11%
Boruta dataset: 78.46%

Support Vector Machine

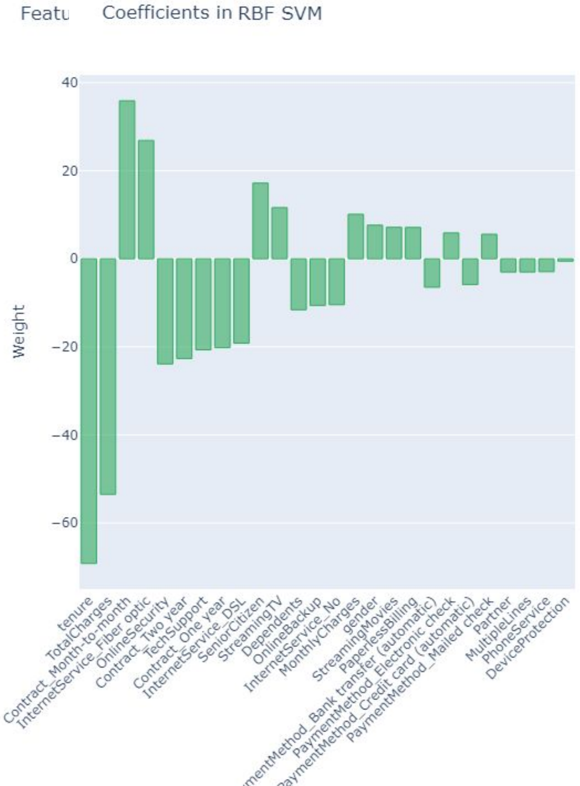
Linear SVM



Polynomial SVM



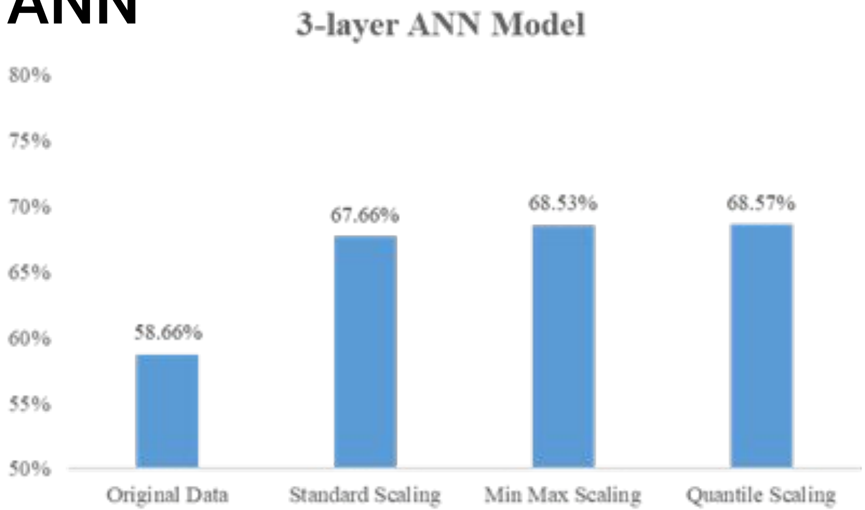
RBF SVM



3.3

Artificial Neural Network (ANN)

ANN

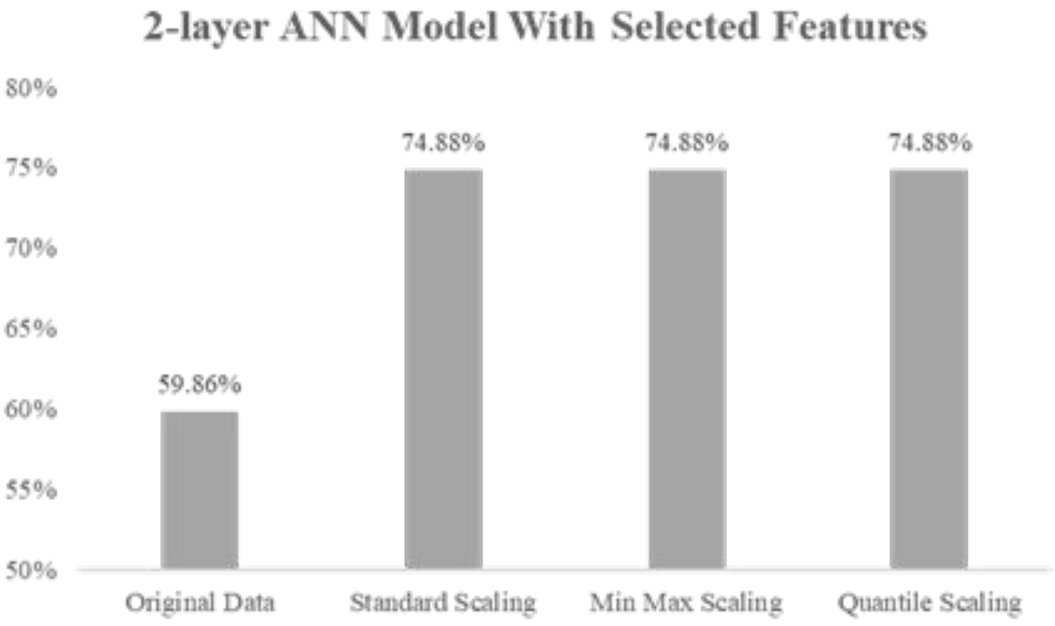


Fit on full dataset

- We consider two- and three-hidden-layer ANN models according to the problem's complexity.
- After scaling, the prediction accuracy has been significantly improved
- Min-max and quantile scaling perform better
- The accuracy of 2-layer model is higher

ANN

Fit on 11 selected features



- The accuracy after scaling increases slightly, comparing to full model.
- The accuracy does not change after we reduce the node's number and epochs.
- The overall performance of the ANN model are not as good as other simpler models

3.4

XGBoost

What is XGBoost

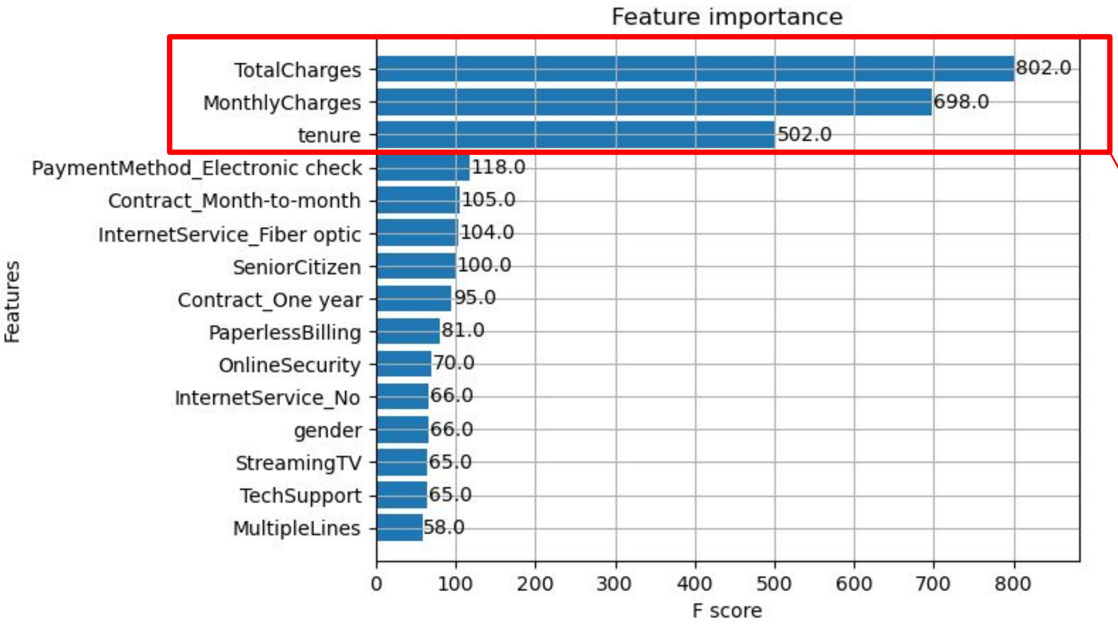
How it works

- XGBoost improves its prediction accuracy by integrating weak decision tree classifiers.
- XGBoost uses second-order gradient information as the loss function

Advantage

- XGBoost will perform regularization during the training process which can avoid the huge influence of outliers

XGBoost

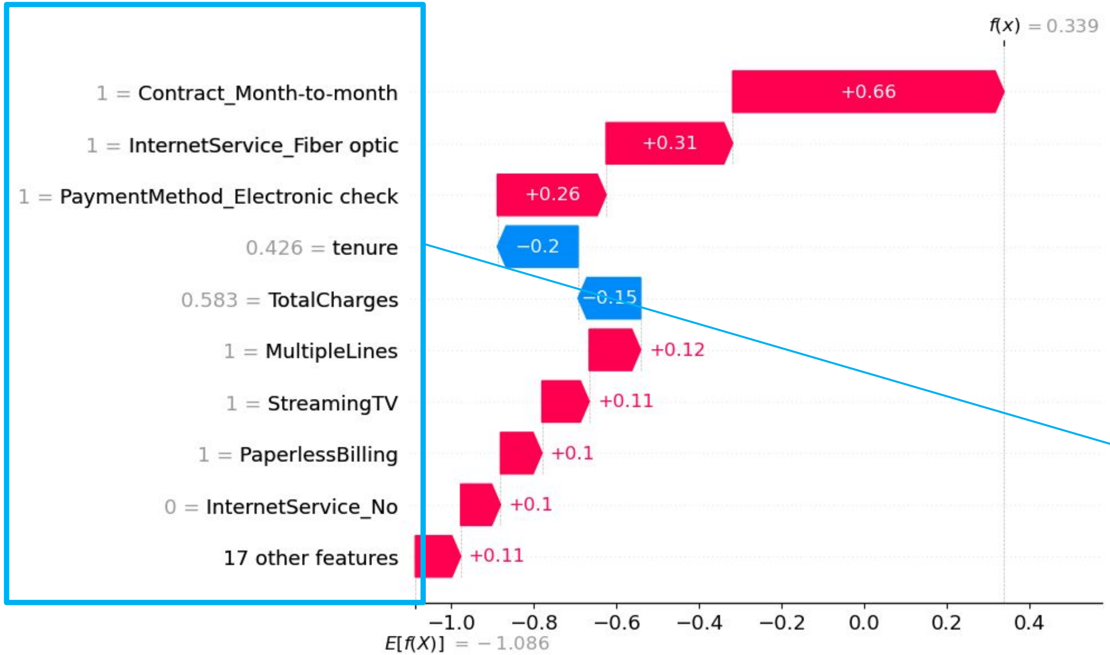


Fit with original dataset

- Test accuracy is 81.37%, best among all other models
- Important variables are Total Charges, Monthly Charge and tenure.

XGBoost

SHAP diagram

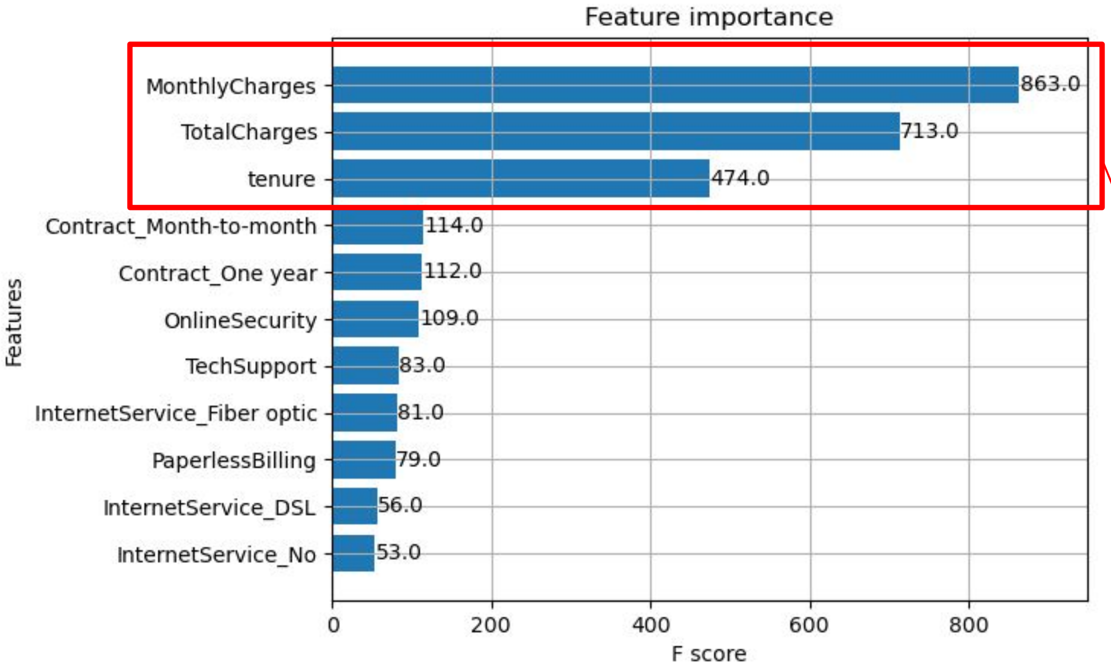


No Monthly_Charges

Fit with original dataset

- SHAP diagram is used to compare and see the contribution strength of different variables to the final prediction result.
- the Monthly Charges with high feature importance in previous slides did not appear in the SHAP graph

XGBoost

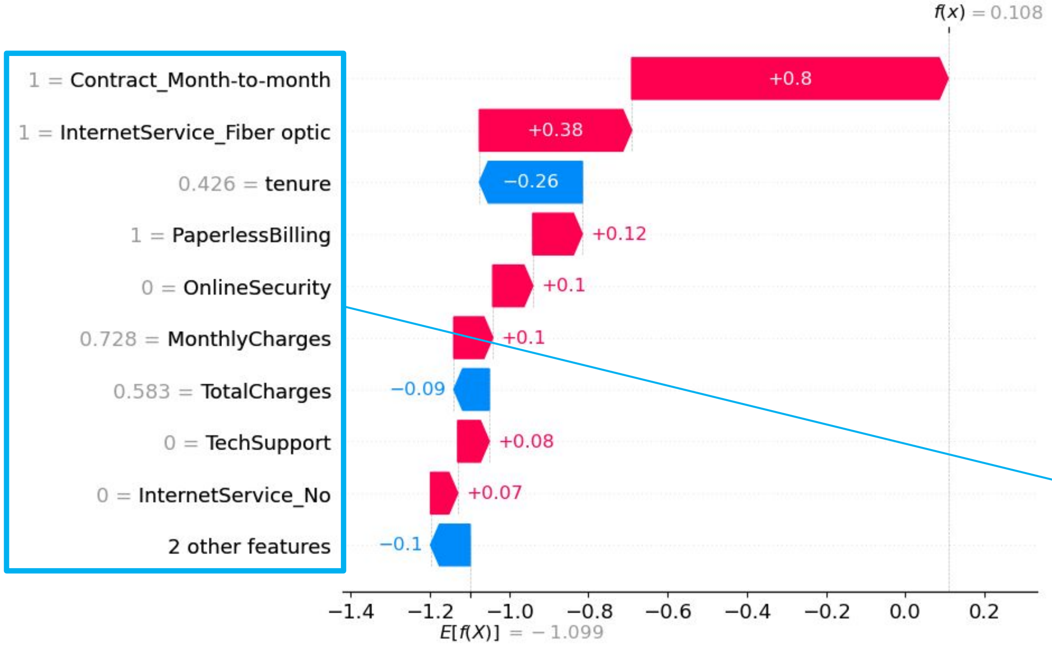


Fit with 11 selected features

- Test accuracy is 80.81%, slight lower than full model
- Important variables are still Total Charges, Monthly Charge and tenure.

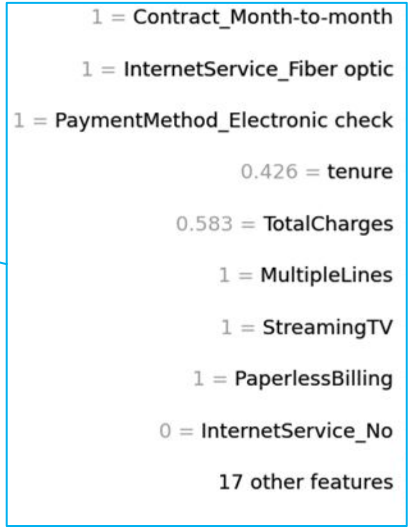
XGBoost

SHAP diagram



Fit with 11 selected features

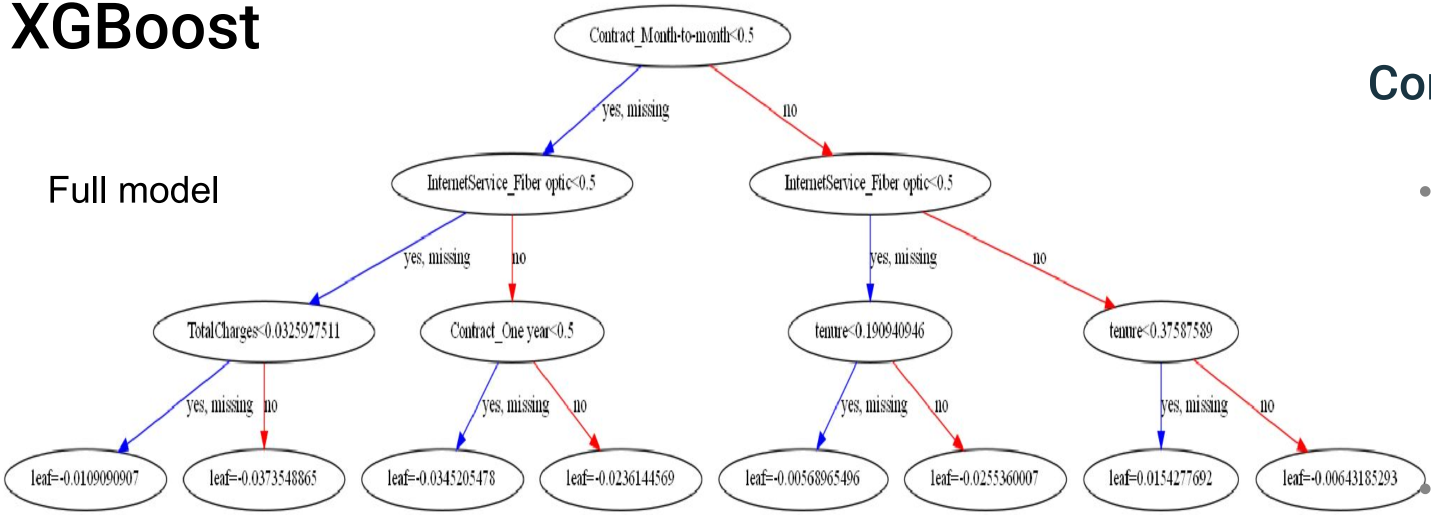
- Substantial changes in SHAP diagram
- Indicate that some variables are indeed unnecessary



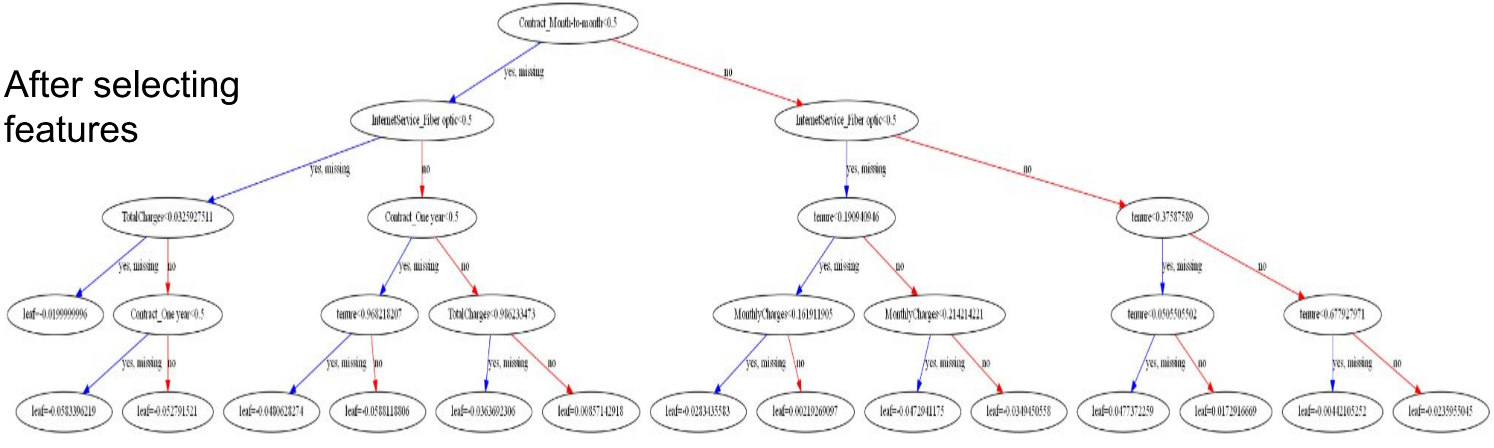
Previous SHAP plot

XGBoost

Full model



After selecting features



Comparing tree diagram

- model tree after selecting features is more complex and deeper,
- There may be an
erfitting problem.

e still choose the full
riable model

4. Conclusion

Summary and recommendation

Best Model

XGBoost

- fitted on original scaled dataset

Reasons for selection

1. Highest accuracy of all models at 81.37%
2. Provides additional properties like regularization
3. Gives a resulting tree with relative small depth

Reasons against feature selection

Feature selection does not perform well due to the factors:

- Small sample size (as a percentage of total telecommunication industry)
- High dimensionality

This is supported by the following papers:

- **Zheng et al. (2007)** → In datasets with many features, some features may have low discriminatory power, difficult to identify the most relevant features
- **Meinshausen and Bühlmann (2010)** → Performance of feature selection methods are dependent on sample size, and sometimes not reliable if sample size is too small

References

Bengio, Y., Goodfellow, I., & Courville, A. (2016). Deep learning. MIT Press.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284
https://doi.org/10.1109/TKDE.2008.239

Liuzhi Yin, Yong Ge, Keli Xiao, Xuehua Wang, Xiaojun Quan,Feature selection for high-dimensional imbalanced data, Neurocomputing, Volume 105, 2013, Pages 3-11, ISSN 0925-2312,
https://doi.org/10.1016/j.neucom.2012.04.039

Zheng, Z., Wu, X., & Srihari, R. K. (2007). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-07) (pp. 856-863).

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417-473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Thank you!

Does anyone have any questions?