Customer Retention Strategy Recommendation for Telecom Companies

Carol Yin
Rose Ro
Cece Li
Kailin Yu
Jack Gao

Agenda

1. Project Overview
2. Exploratory Data Analysis
3. Modeling Methodology
4. Findings & Recommendations

1.

Project Overview

A Changing Landscape

Nowadays, telecommunication industries are extremely saturated, and service providers are experiencing high rates of churn

Rising expectations from customers

With the transition to 5G services, customers' expectation for quality of services are higher than ever. However, industry is extremely saturated, and customers have an abundance of choices.

Quicker switch and shorter tenure

According to a report by Accenture[1], **77%** of customers switch their telecom providers more quickly than they did 3 years ago.

Greater financial benefit from customer retention

A report from Bain & Company[2] indicates that a **5%** increase in a company's retention rate would boost profits by up to **95%**.

Source:

1. https://www.accenture.com/t20170216T035010Z__w__/us-en/_acnmedia/PDF-43/Accenture-Strategy-GCPR-Customer-Loyalty.pdf#zoom=50
2. https://cloudcherry.com/blog/improve-customer-retention-in-telecom/

How We Can Help

Leverage data mining and machine learning to help telecom companies improve customer retention rate

Provide Telecom Companies with

1. Classification Models

Based on the customer data obtained, we used various feature engineering techniques and trained several ***classification models*** to select the most ***optimal*** *ones with* ***high predicting power.***

2. Key Factors

Besides providing the optimal models, we also dig deeper and discover the ***key factors*** that would have the ***biggest impact*** on customer's churn rate.

3. Customized Recommendations

In the end, we will provide specific ***business strategies*** that are easy to implement, from both a ***short-term*** and ***long-term*** perspective.

2.
Exploratory Data Analysis

How Raw Dataset Looks Like

Our dataset is clean with no missing values and has a lot of binary columns

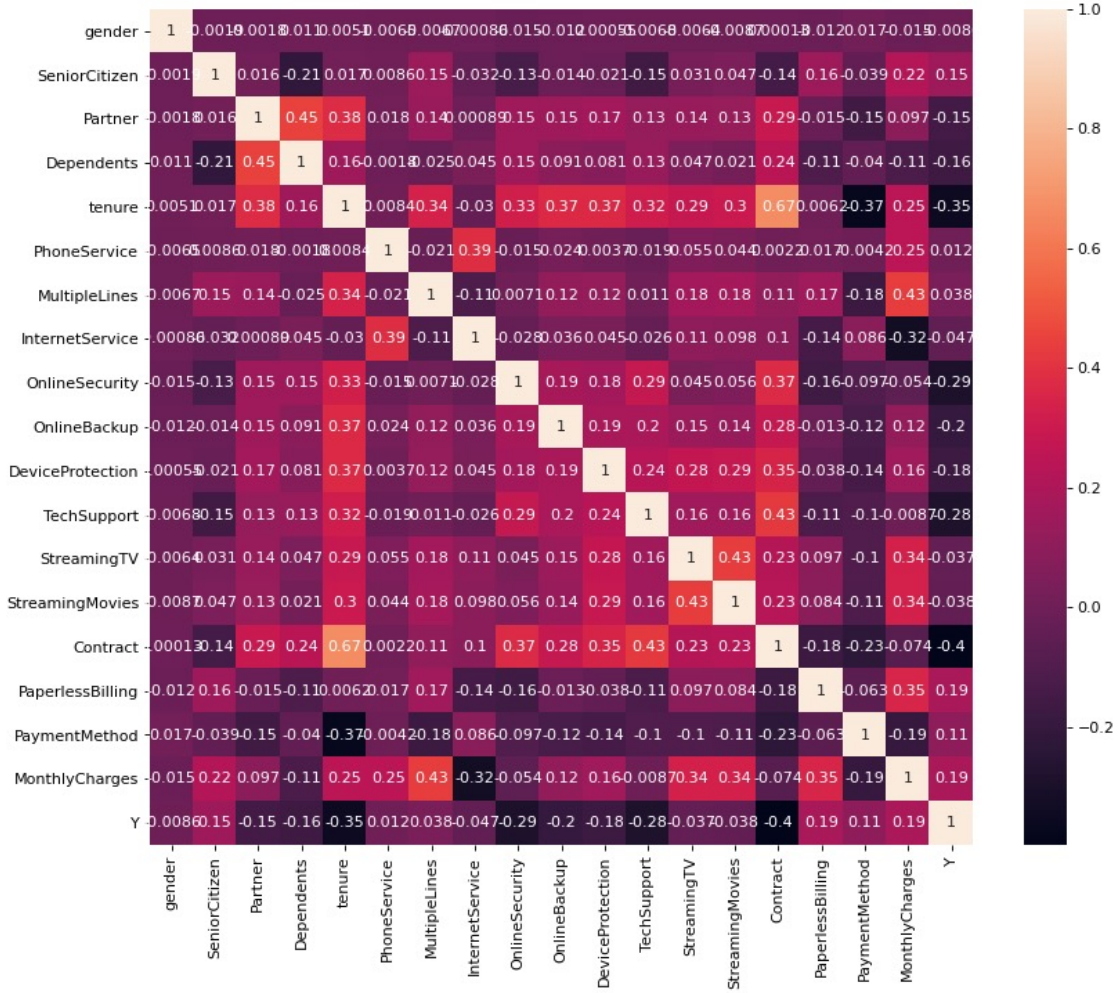
- Data source: Kaggle (Originally from IBM Watson analytics data)
- 7043 rows x 20 columns
- No missing values
- Each row indicates a unique customer who is identified by the column 'customerID'
- Our target column: 'Churn'
- 3 numerical columns:
  'tenure', 'MonthlyCharges', 'TotalCharges'
- 17 categorical/binary columns (see right)

Details of Categorical Columns

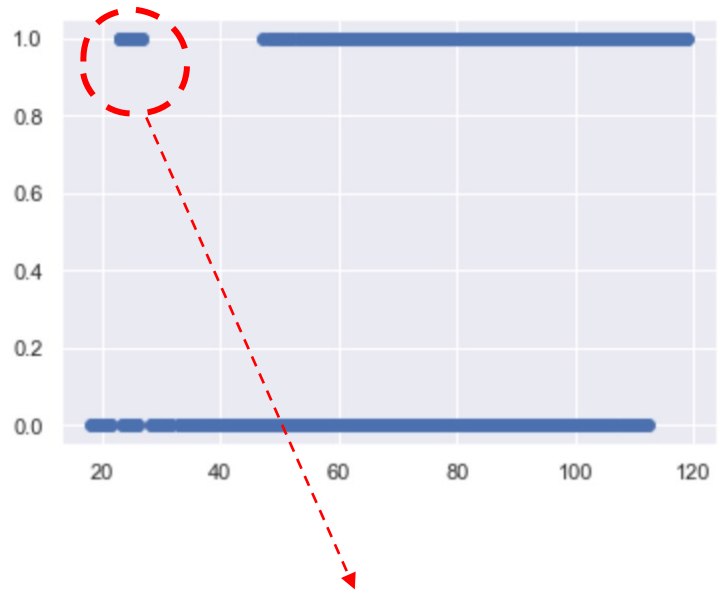
```
gender:  ['Female' 'Male']
SeniorCitizen:  [0 1]
Partner:  ['Yes' 'No']
Dependents:  ['No' 'Yes']
PhoneService:  ['No' 'Yes']
MultipleLines:  ['No phone service' 'No' 'Yes']
InternetService:  ['DSL' 'Fiber optic' 'No']
OnlineSecurity:  ['No' 'Yes' 'No internet service']
OnlineBackup:  ['Yes' 'No' 'No internet service']
DeviceProtection:  ['No' 'Yes' 'No internet service']
TechSupport:  ['No' 'Yes' 'No internet service']
StreamingTV:  ['No' 'Yes' 'No internet service']
StreamingMovies:  ['No' 'Yes' 'No internet service']
Contract:  ['Month-to-month' 'One year' 'Two year']
PaperlessBilling:  ['Yes' 'No']
PaymentMethod:  ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn:  ['No' 'Yes']
```

Detect Feature Collinearity & Outliers

No significant collinearity between features, nor do we have outliers




```
#The points at the top left corners could be considered outliers...?
sns.set()
#cols = ['MonthlyCharges','MultipleLines_Yes']
#sns.pairplot(x1[cols])
plt.scatter(x1['MonthlyCharges'],x1['MultipleLines_Yes'])
plt.show()
```



These were more than 350 points. So not considered outliers

Deal With Data Imbalances

SMOTE with a combination of oversampling and undersampling slightly improved our model

- Initial train dataset churn ratio: Yes : No = 1 : 2.77 (Yes: 1308, No: 3622)
- This ratio Is not too bad: > 1:5 ratio
  - which is also why resampling did not significantly improve our model

Oversampling Experiments with Random Forest as the Base Model

Resampling Method	Accuracy	F1 Score(0/1)
Simple Resampling	0.751	0.81/0.62
SMOTE (over only, regular)	0.760	0.82/0.62
SMOTE (over only, SVM)	0.752	0.82/0.62
SMOTE (over only, Borderline)	0.750	0.82/0.61
★ SMOTE (over with regular and under)	0.772	0.84/0.62
SMOTE (over with Borderline and under)	0.763	0.83/0.62
ADASYN	0.753	0.82/0.62
Without Resampling	0.805	0.87/0.57

The best resampling method
oversampled the minority class to 3:5 ratio and undersampled the majority class to 4:5 ratio

```
over = SMOTE(sampling_strategy=0.6)
under = RandomUnderSampler(sampling_strategy=0.8)
```

Reference:
https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

3.

Modeling Methodology

Try Different Variable Transformations

Get-Dummies

gender_Male	Partner_Yes	Dependents_Yes	PhoneService_Yes	MultipleLines_No phone service	MultipleLines_Yes
0	1	0	0	1	0
1	0	0	1	0	0
1	0	0	1	0	0
1	0	0	0	1	0
0	0	0	1	0	0

Label Encoding

gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
0	0	1	0	1	0	2
1	0	0	0	34	1	0
1	0	0	0	2	1	0
1	0	0	0	45	0	2
0	0	0	0	2	1	0

Model	Accuracy	F1 Score(0/1)
Random Forest	0.805	0.87/0.57
KNN	0.791	0.86/0.54
★ Logistic	0.811	0.88/0.61
SVM	0.789	0.86/0.55
Decision Tree	0.798	0.87/0.55

≈

Model	Accuracy	F1 Score(0/1)
Random Forest	0.804	0.88/0.52
KNN	0.800	0.87/0.60
Logistic	0.808	0.87/0.60
SVM	0.807	0.88/0.57
Decision Tree	0.801	0.87/0.57

Observations:

- Get-dummies method along with Logistic regression model reaches highest accuracy and F1 Score.

Assign Two Grouping methods

- ***Combine***: treat 'No phone service' 'No internet service' all as 'No' (0)
- ***Non-Combine***: keep the original group

```
data['MultipleLines'].replace({'Yes' : 1, 'No' : 0, 'No phone service': 0})
data['OnlineSecurity'].replace({'Yes' : 1, 'No' : 0, 'No internet service': 0})
data['OnlineBackup'].replace({'Yes' : 1, 'No' : 0, 'No internet service': 0})
data['DeviceProtection'].replace({'Yes' : 1, 'No' : 0, 'No internet service': 0})
data['TechSupport'].replace({'Yes' : 1, 'No' : 0, 'No internet service': 0})
data['StreamingTV'].replace({'Yes' : 1, 'No' : 0, 'No internet service': 0})
data['StreamingMovies'].replace({'Yes' : 1, 'No' : 0, 'No internet service': 0})
```

Performance with **Combine**

Model	Accuracy	F1 Score(0/1)
Random Forest	0.805	0.87/0.57
KNN	0.791	0.86/0.54
Logistic	0.811	0.88/0.61
SVM	0.789	0.86/0.55
Decision Tree	0.798	0.87/0.55

>

Performance with **Non-Combine**

Model	Accuracy	F1 Score(0/1)
Random Forest	0.803	0.87/0.56
KNN	0.788	0.86/0.53
Logistic	0.81	0.88/0.61
SVM	0.788	0.86/0.55
Decision Tree	0.799	0.87/0.55

Observations:

- The accuracy with each model is obviously higher when using binary class.
- Model accuracy not only depends on model selection but also data grouping methods.

Apply Dimension Reduction Technique

Performance **without PCA**

Model	Accuracy	F1 Score(0/1)
Random Forest	0.804	0.88/0.52
KNN	0.80	0.87/0.60
Logistic	0.81	0.87/0.60
SVM	0.808	0.88/0.57
Decision Tree	0.802	0.87/0.57

Performance **with PCA**

Model	Accuracy	F1 Score(0/1)
Random Forest	0.806	0.88/0.55
KNN	0.81	0.87/0.62
Logistic	0.81	0.88/0.60
SVM	0.813	0.88/0.59
Decision Tree	0.805	0.88/0.56

Observations:

- There is no considerable improvement in the performance of each model after applying PCA
- After doing some research, we found that PCA is more useful on continuous numeric data rather than data with many binary features

Add More Features From Twitter

Our Goal:

- Add polarity (Positive vs.. Negative) values extracted from twitter text using sentimental analysis to original data as an extra feature, to bring more information to the data.

Steps:

1. Extract polarity value from twitter post by using sentimental analysis, higher values means user's positive attitudes and vice versa. (See Figure 1)
2. Cluster original data to 4 groups using K-Means and decide the probability of Churn within each group by observing key features and metrics. (See Figure 2)
3. Assign the polarity value to each group according to high likely to churn → not likely to churn order. Resampling the data for future model training.

Tweet	Polarity
Making sure you get the most out of T-Mobile T...	0.445312
The Note was just launched. I'm so glad that I...	0.166667
I get penalized for being broke with my T-Mob...	0.000000
Just changed my Sprint SIM card for a T-mobile...	-0.015152
Oh I hear you! Between my wife and me we are ...	0.000000

Figure 1

	x_0	x_1	x_2	x_3
tenure	10.594842	58.583461	14.766697	54.169124
MonthlyCharges	32.643123	93.283002	81.085178	33.964224
PhoneService_Yes	0.797135	0.988259	0.993596	0.747615
MultipleLines_Yes	0.097994	0.761103	0.463861	0.255854

Figure 2

Selected Classification Models

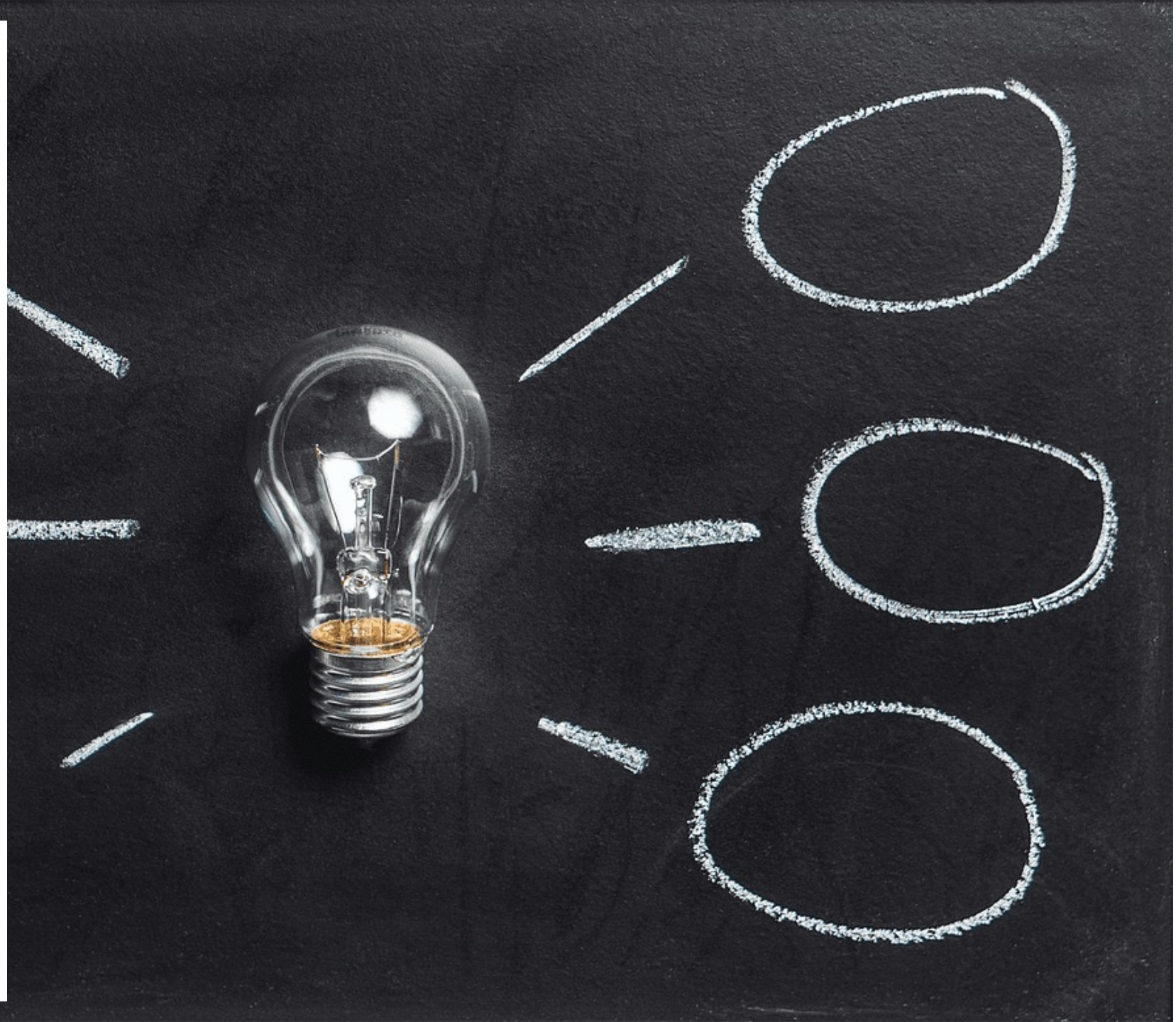
	Advantages	Disadvantages
Logistic Regression	Easy to implement, fast	Only applicable for linear relationship, may suffer from underfitting
KNN	Easy to implement, very fast	Do not work well with large dataset with high dimension, sensitive to noise
SVM	Effective in data with high dimension, memory efficient	Not suitable for large dataset, may suffer from overfitting
Decision Tree	Easy to understand, low requirement for dataset	May suffer from overfitting, time inefficiency
Random Forest	Tend to not overfit, can extract feature importance	Less interpretable than single decision tree, require significant memory for storage

Final Decisions for Modeling

- *Resample training data using **SMOTE** (hybrid of regular and under)*
- *Choose **Get Dummies** and **Combine** for data transformation*
- *Use augmented data with **tweet sentiments***

4.

Findings & Recommendations



Final Model Comparison

Figure 1 Results of F1 Score

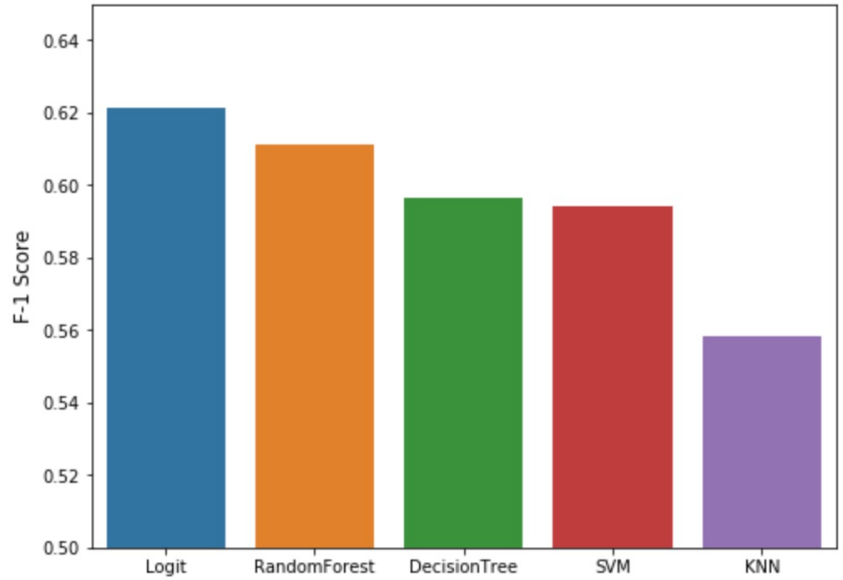


Figure 2 Final Model Results

	Accuracy	Precision	Recall	F-1
Logit	0.791292	0.599338	0.645276	0.621459
RandomForest	0.782773	0.582258	0.643494	0.611346
DecisionTree	0.758637	0.536273	0.672014	0.596519
SVM	0.758164	0.535817	0.666667	0.594122
KNN	0.726455	0.488621	0.650624	0.558104

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$$

Metric Choice

- *F1 Score* is the harmonic mean of the ***precision*** and ***recall***, which gives a better measure of the incorrectly classifies cases and penalizes extreme values.
- In most real-life classification problems where imbalanced class distribution exists, using the ***F1 Score*** would be a better metric than merely using the ***accuracy***.

Analysis of Key Factors

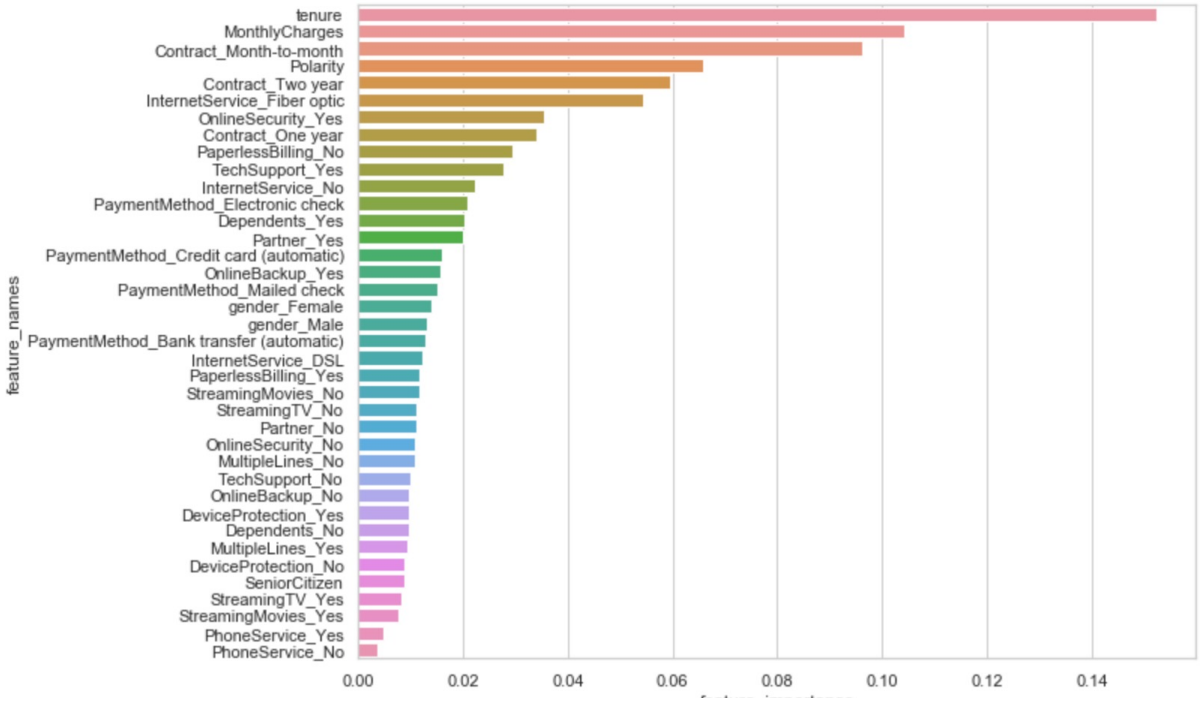


Figure 1 Feature Importance of the *Random Forest* Model

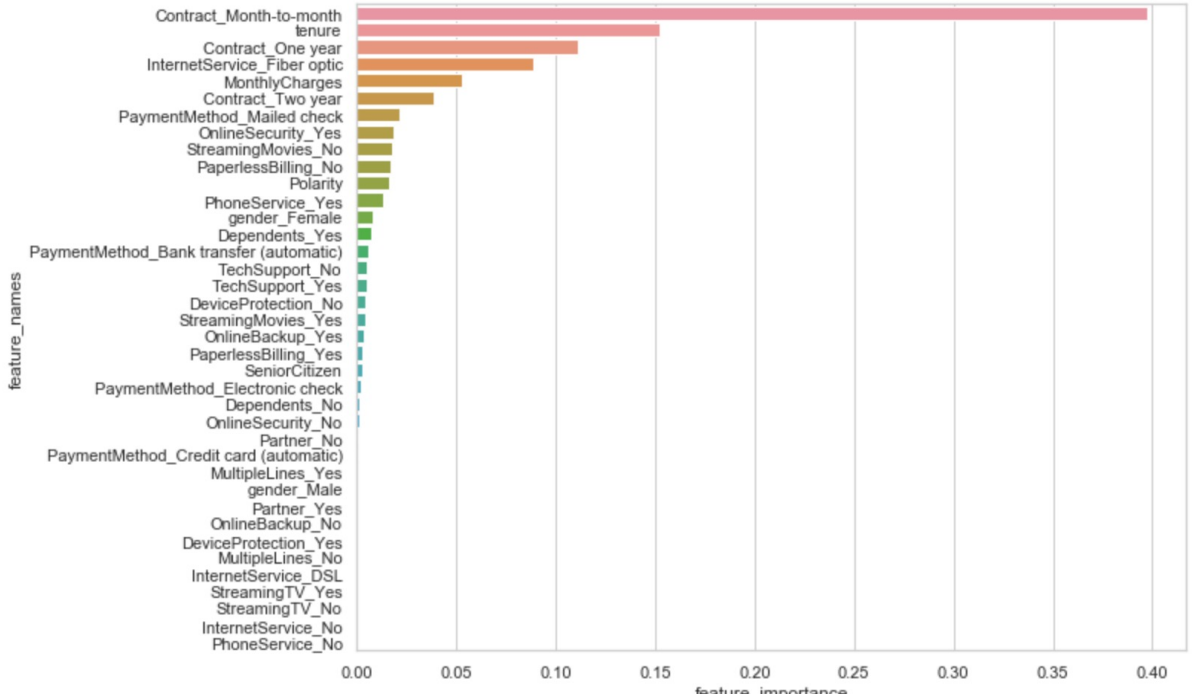


Figure 2 Feature Importance of the *Decision Tree* Model

Top 5 Features

1. Tenure	1. Contract Month to Month
2. Monthly Charges	2. Tenure
3. Contract Month to Month	3. Contract One Year
4. Polarity	4. Internet Service (Fiber Optic)
5. Contract Two Year	5. Monthly Charges

Business Strategies to Recommend

1 ***Offer Long-Term Contracts if Possible***

- Offering a longer subscription model can extend customer commitment, giving them more time to use and see the benefit of the product.

2 ***Provide Discounts and Incentives to High-Risk Customers***

- Offer incentives, such as discounts and special offers to customers who are identified as likely to defect. But be sure the cost of increasing customer retention rate do not outweigh the profit to be gained from saved customers.

3 ***Be Proactive and Responsive to Customer's Complains and Questions on Social Platforms***

- Younger generations are active and tend to express their opinions and feelings on social platforms. It is key for telecom providers to address their requests in a proper and timely manner.

4 ***Invest in Innovation and Create Competitive Advantages Over Others***

- Innovation has been and always will be the differentiator and game changer in the telecom industry. Costumers are less likely to leave providers with cutting-edge technologies that are just different from others.


Questions?