Determining Telco Churning.

By RMM - July 2018

Mid-term Capstone.

This project corresponds to the final part of Unit 3 at the mid time of the Thinkful Data Science Bootcamp. You can visit this notebook at my GitHub repository.

Outline

- Introduction
- Dataset information & modelling aspects
- Feature & variable analysis
- Analysing Dataset
- Predicting Churn
- Most important features and variables in customer churn
- Probability of customers leaving
- How can this information help us?
- How could we optimize our conclusions?

Intro

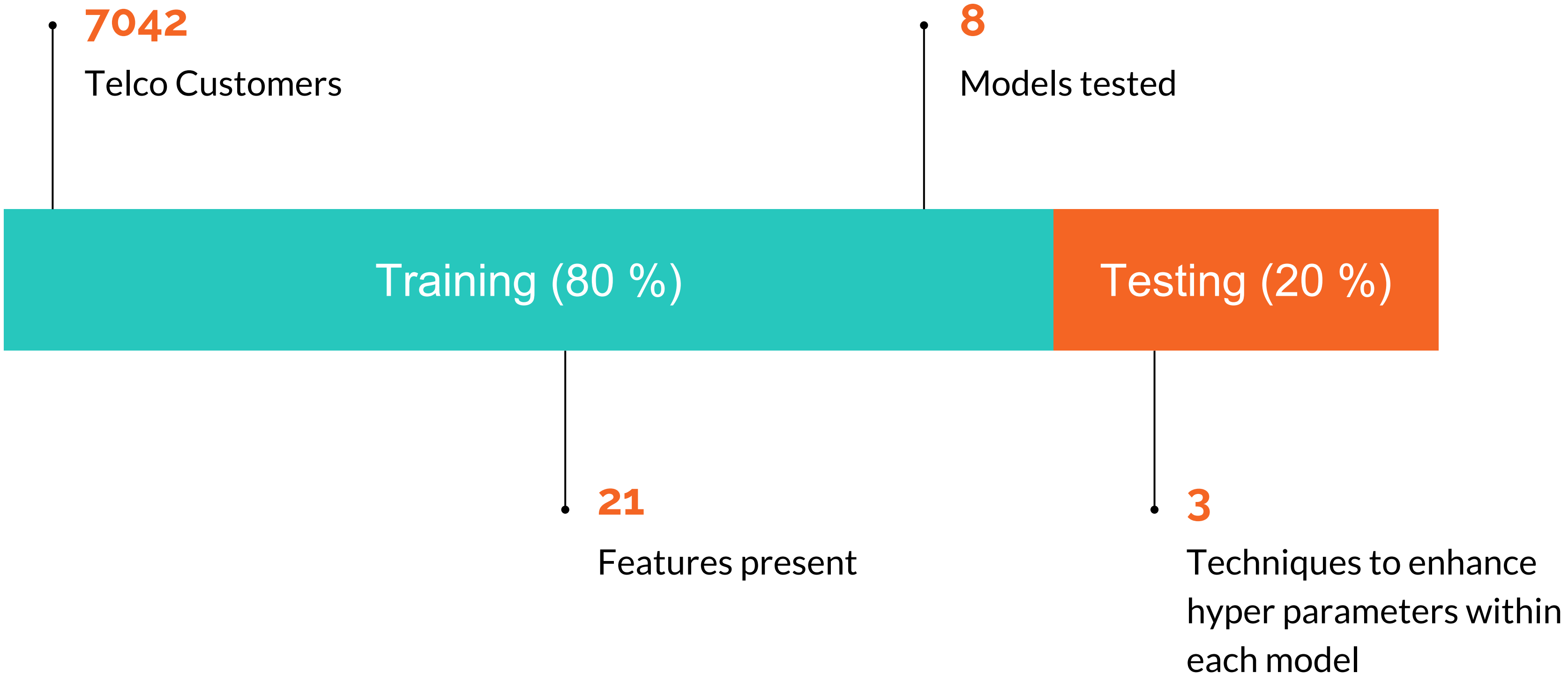
Customer churn (cc) occurs when customers or subscribers stop doing business with a company or service.

➔ **Which features influence cc?**
Valuable for building a retention campaign

➔ **Which features are most important?**
Knowing this will allow us to reduce computational costs and focus our resources when building our retention campaign

➔ **Which clients are most likely to leave?**
Applying our best model we'll determine which clients are more likely to leave.

Telco customer Churn - IBM Watson Analytics community



7042

Telco Customers

8

Models tested

Training (80 %)

Testing (20 %)

21

Features present

3

Techniques to enhance hyper parameters within each model

Analysing Dataset.

Month-to-month contracts = 10 month median tenure

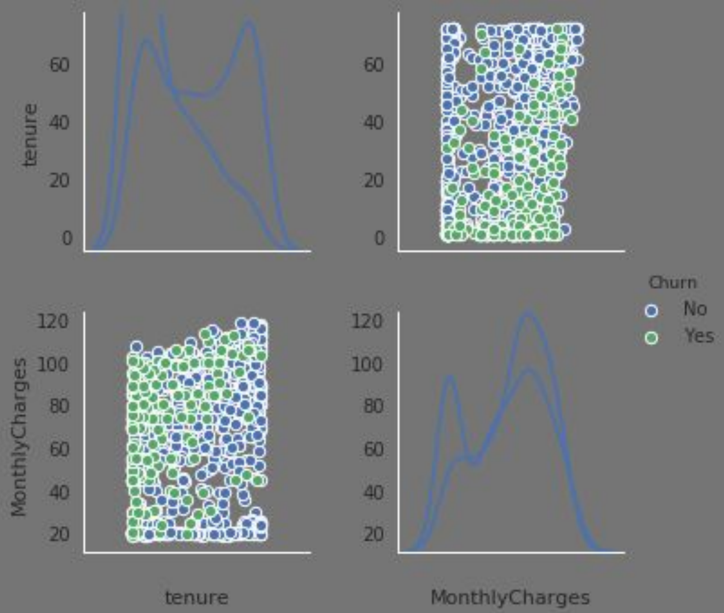
Two year contracts = 70 month median tenure

Tip

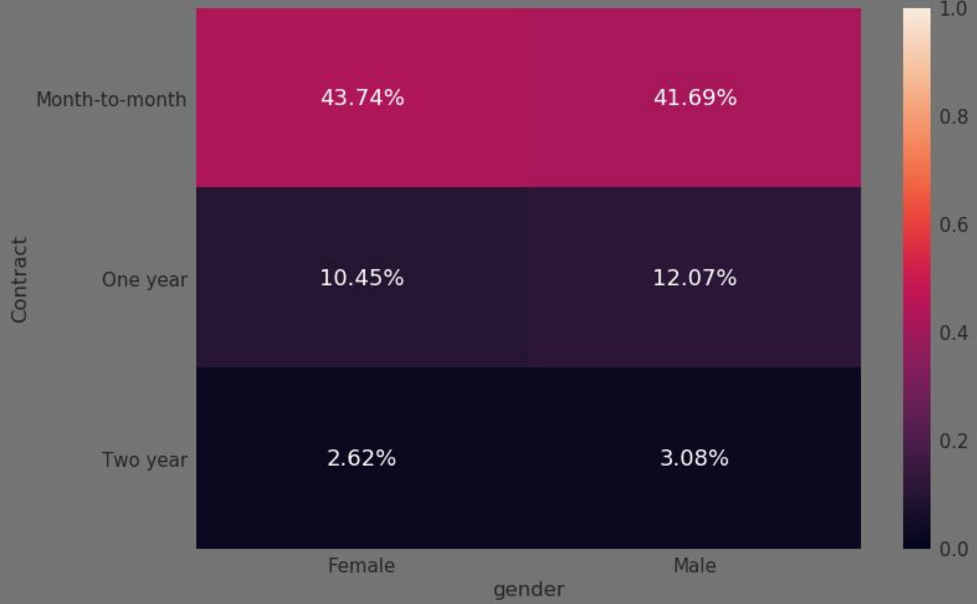
Tenure is the measurement of the amount of time a person is your customer, or in other words it's the age of a customer in your system.

Customers with online backup have a median age in the company of 32 months more in contrast to those without this service

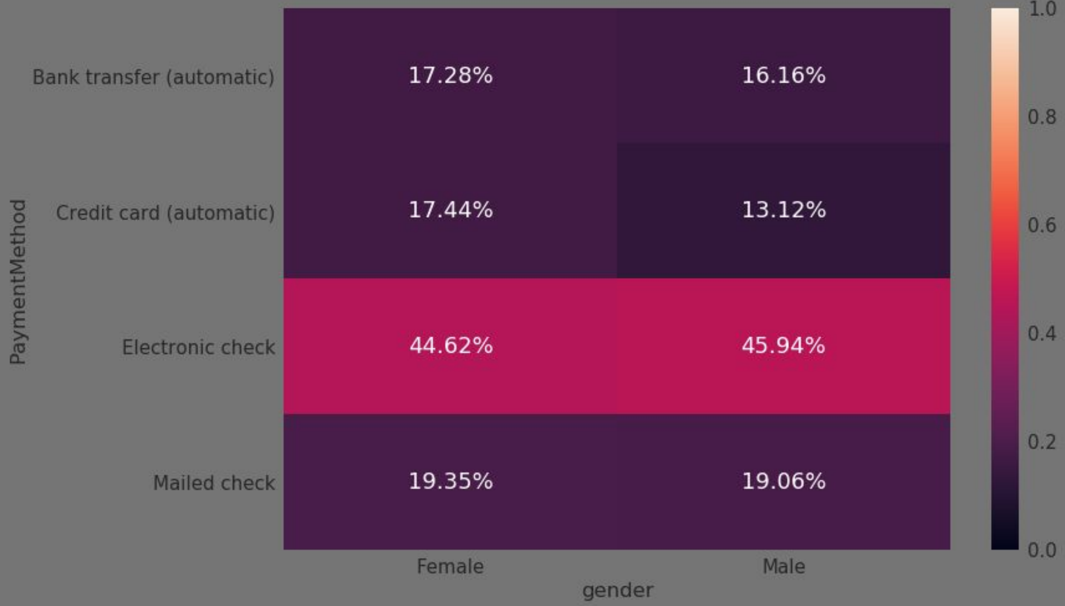
Monthly charges and tenure share a close relationship



Month-to-month contracts have a higher churn probability



Electronic check payers churn most often



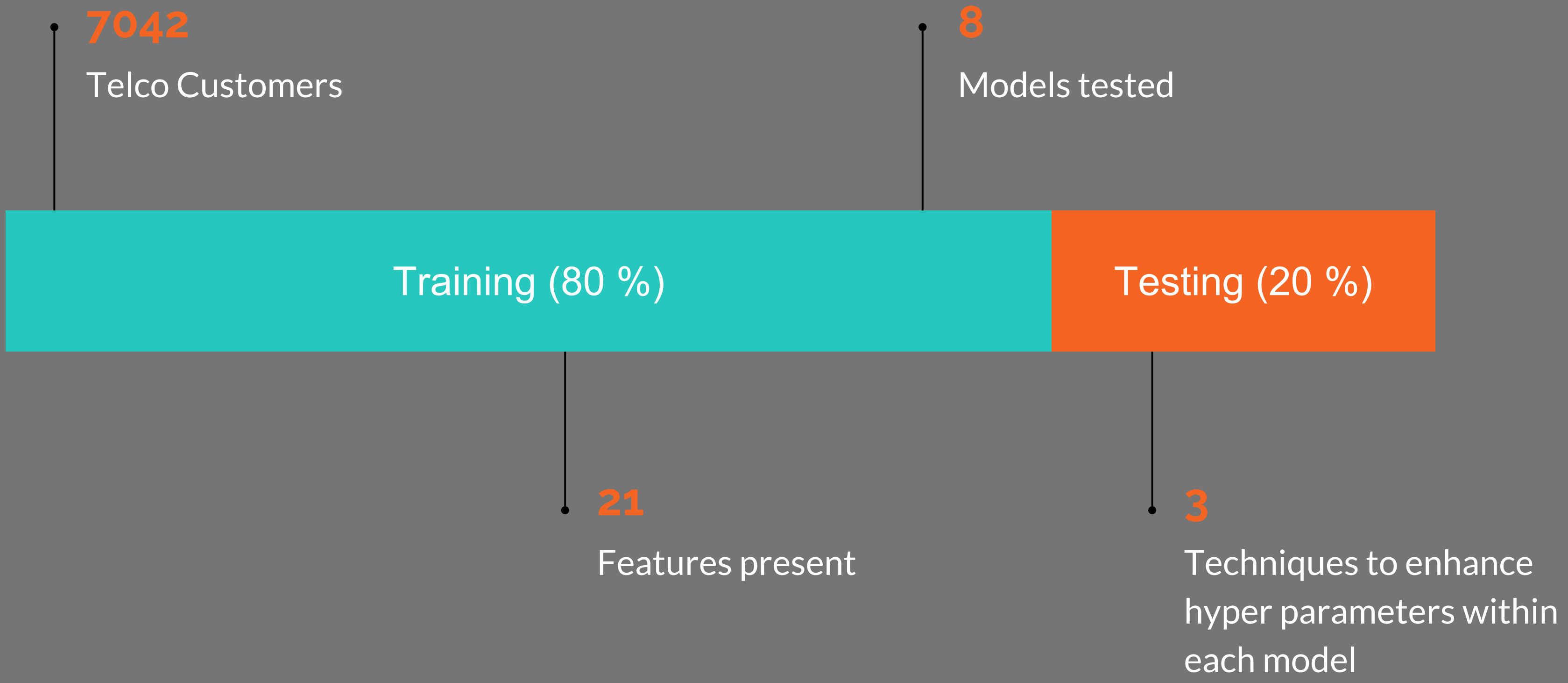
Most influential features

>>

- Tenure
- Monthly charges
- Internet service
- Online security
- Online backup
- Tech support
- Contract
- Payment method

Predicting Churn.

Telco customer Churn - IBM Watson Analytics community



7042

Telco Customers

8

Models tested

Training (80 %)

Testing (20 %)

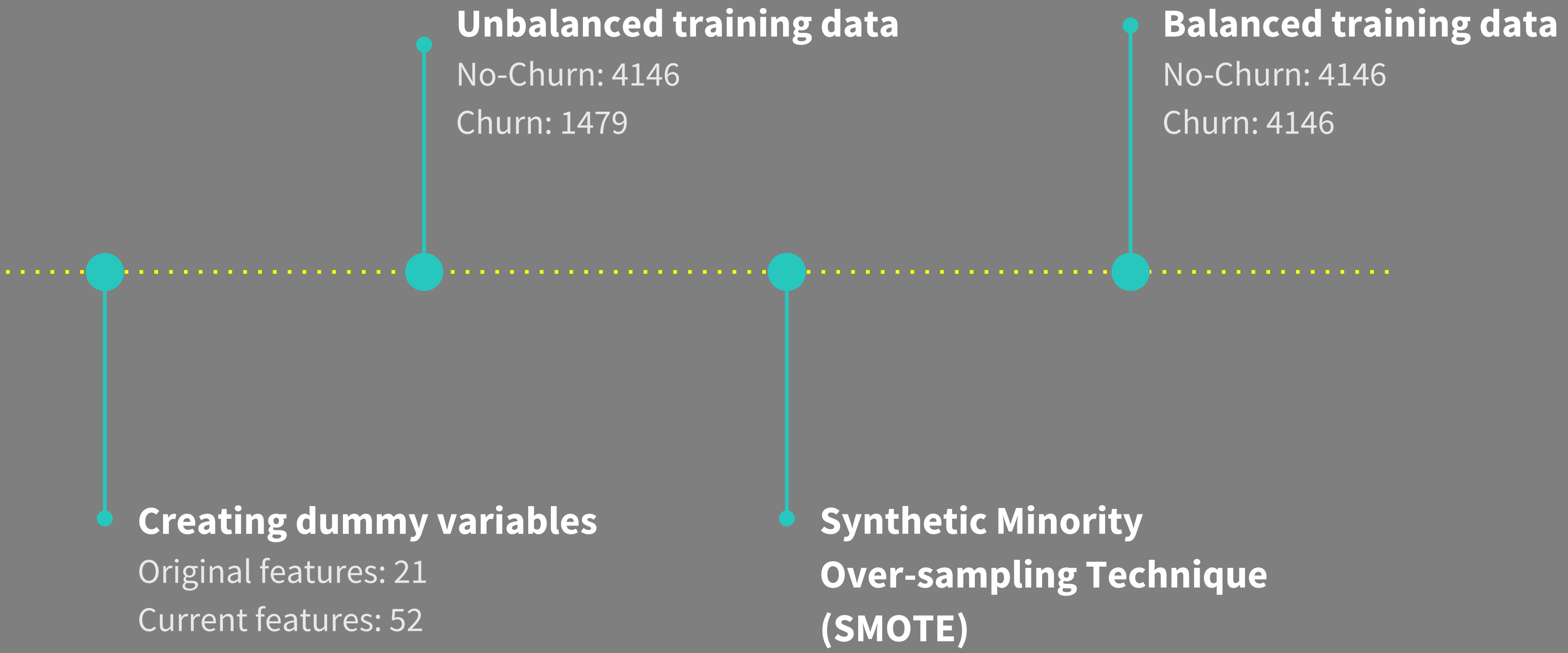
21

Features present

3

Techniques to enhance hyper parameters within each model

Creating dummy variables and balancing the dataset.



Creating dummy variables

Original features: 21

Current features: 52

Unbalanced training data

No-Churn: 4146

Churn: 1479

Synthetic Minority Over-sampling Technique (SMOTE)

Balanced training data

No-Churn: 4146

Churn: 4146

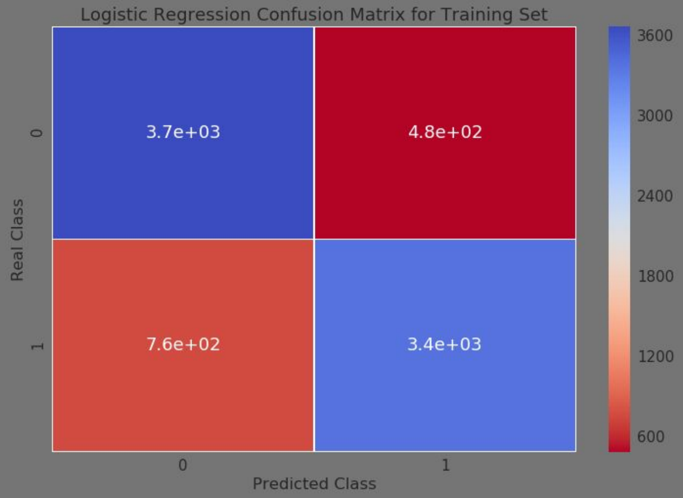
Hyperparameter tuning: GridSearchCV

Logistic regression (Classification L2) classification report:

	precision	recall	f1-score
Churn_No	0.83	0.88	0.86
Churn_Yes	0.88	0.82	0.85
avg / total	0.85	0.85	0.85

Cross Validation Accuracy Scores: 0.840(+/- 0.18)

Cross Validation Accuracy Scores - Test Set: 0.785(+/- 0.05)



grid.best_score_ : 0.84008683068
grid.best_params_: {'C': 1}

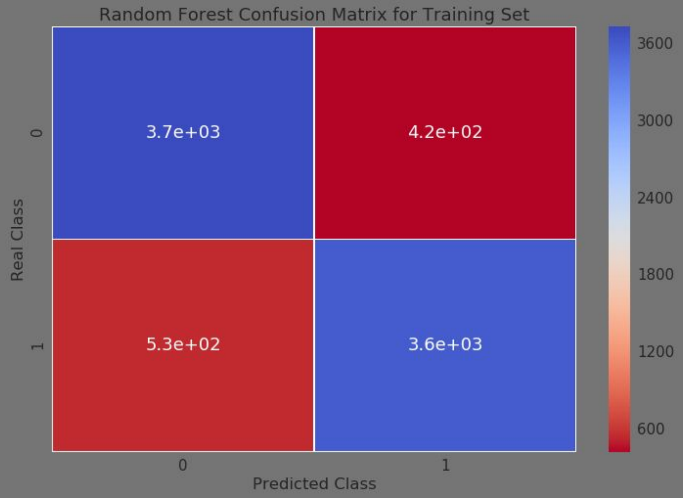
Hyperparameter tuning: GridSearchCV

Random Forest Classification classification report:

	precision	recall	f1-score
Churn_No	0.87	0.90	0.89
Churn_Yes	0.90	0.87	0.88
avg / total	0.89	0.89	0.89

Cross Validation Accuracy Scores: 0.792(+/- 0.15)

Cross Validation Accuracy Scores - Test Set: 0.741(+/- 0.07)



grid.best_score_ : 0.801616015437
grid.best_params_: {'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 1}

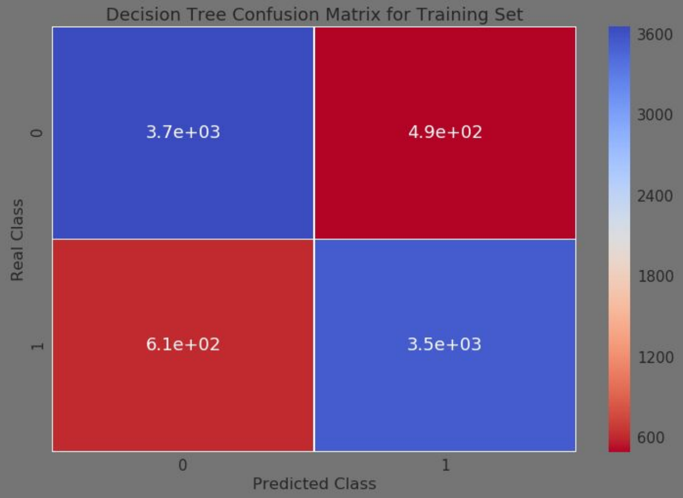
Hyperparameter tuning: GridSearchCV

Decision Tree Classifier classification report:

	precision	recall	f1-score
Churn_No	0.86	0.88	0.87
Churn_Yes	0.88	0.85	0.87
avg / total	0.87	0.87	0.87

Cross Validation Accuracy Scores: 0.809(+/- 0.15)

Cross Validation Accuracy Scores - Test Set: 0.738(+/- 0.09)



grid.best_score_ : 0.812952243126
grid.best_params_ : {'criterion': 'gini', 'max_depth': 13, 'min_samples_split': 10}

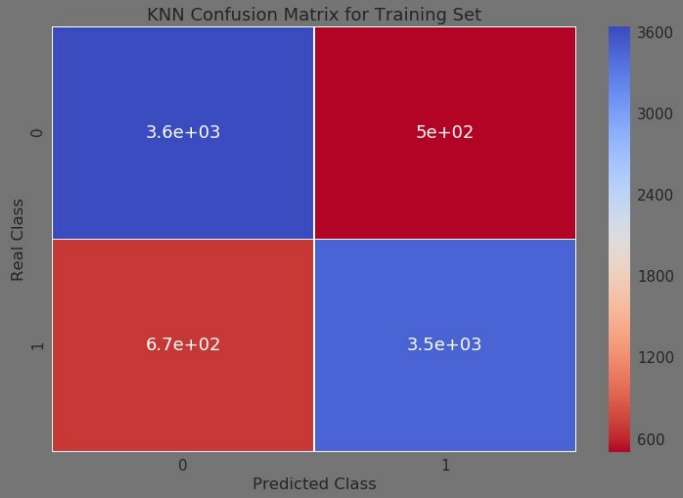
Hyperparameter tuning: GridSearchCV

K-Nearest Neighbours Classifier classification report:

	precision	recall	f1-score
Churn_No	0.84	0.88	0.86
Churn_Yes	0.87	0.84	0.86
avg / total	0.86	0.86	0.86

Cross Validation Accuracy Scores: 0.826(+/- 0.17)

Cross Validation Accuracy Scores - Test Set: 0.773(+/- 0.06)



grid.best_score_ : 0.826338639653
grid.best_params_: {'n_neighbors': 9}

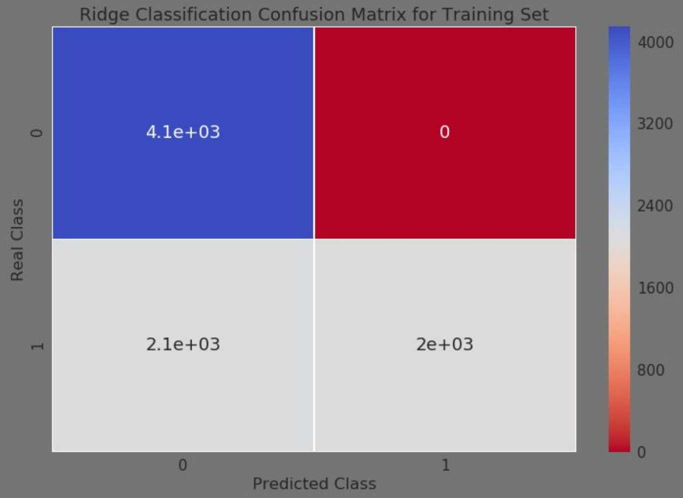
Hyperparameter tuning: GridSearchCV

Ridge Classifier classification report:

	precision	recall	f1-score
Churn_No	0.66	1.00	0.80
Churn_Yes	1.00	0.66	0.66
avg / total	0.83	0.75	0.73

Cross Validation Accuracy Scores: 0.749(+/- 0.36)

Cross Validation Accuracy Scores - Test Set: 0.790(+/- 0.05)



grid.best_score_ : 0.749155812832
grid.best_params_: {'alpha': 0.8}

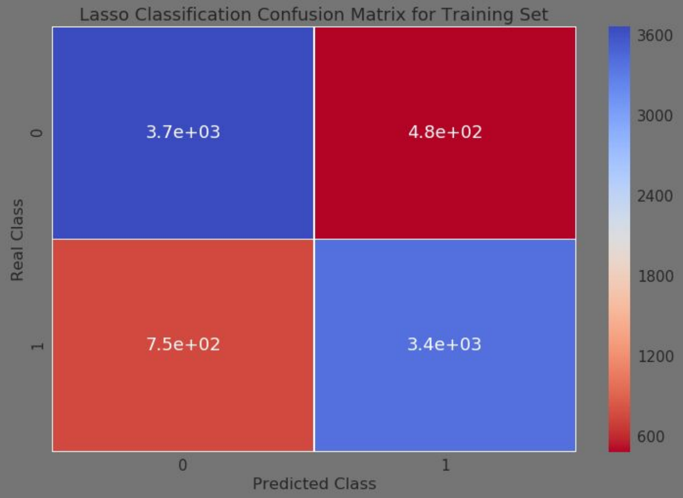
Hyperparameter tuning: GridSearchCV

Lasso Classifier classification report:

	precision	recall	f1-score
Churn_No	0.83	0.88	0.86
Churn_Yes	0.87	0.82	0.85
avg / total	0.85	0.85	0.85

Cross Validation Accuracy Scores: 0.841(+/- 0.18)

Cross Validation Accuracy Scores - Test Set: 0.788(+/- 0.05)



grid.best_score_ : 0.840448625181
grid.best_params_: {'C': 1}

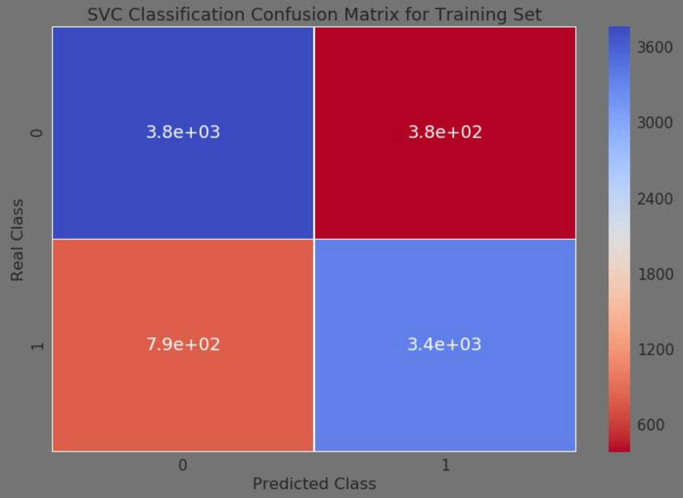
Hyperparameter tuning: GridSearchCV

SVC Classifier classification report:

	precision	recall	f1-score
Churn_No	0.83	0.91	0.86
Churn_Yes	0.90	0.81	0.85
avg / total	0.86	0.86	0.86

Cross Validation Accuracy Scores: 0.837(+/- 0.21)

Cross Validation Accuracy Scores - Test Set: 0.788(+/- 0.05)



grid.best_score_ : 0.836468885673
grid.best_params_: {'C': 10}

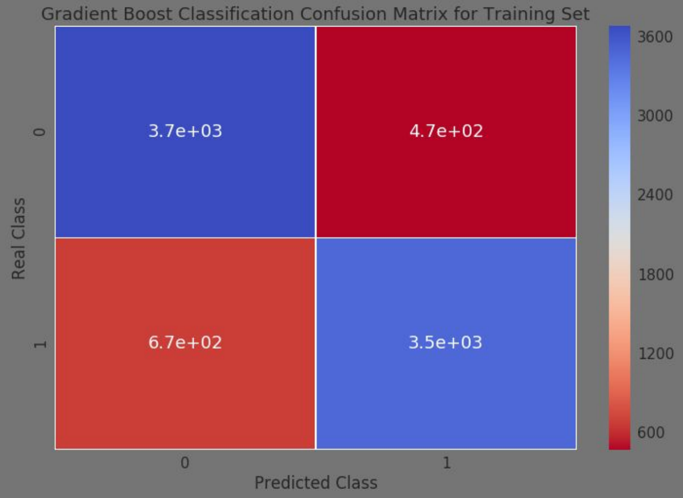
Hyperparameter tuning: GridSearchCV

Gradient Boost Classifier classification report:

	precision	recall	f1-score
Churn_No	0.85	0.89	0.87
Churn_Yes	0.88	0.84	0.86
avg / total	0.86	0.86	0.86

Cross Validation Accuracy Scores: 0.844(+/- 0.19)

Cross Validation Accuracy Scores - Test Set: 0.770(+/- 0.05)



grid.best_score_ : 0.8451519536903039
grid.best_params_: {'max_features': 1.0, 'learning_rate': 0.05, 'max_depth': 4, 'min_samples_leaf': 20}

Predicting chun using PCA.

52 to 38

Fitting PCA to the training matrix, and retaining 90 % of it's variance we reduced the number of features used from 52 to 38 and optimizing prediction score and computational time.

Applying PCA

Logistic regression (Classification L2) classification report:

	precision	recall	f1-score
Churn_No	0.72	0.97	0.83
Churn_Yes	0.95	0.63	0.76
avg / total	0.84	0.80	0.79

Cross Validation Accuracy Scores: 0.780(+/- 0.40)

Cross Validation Accuracy Scores - Test Set: 0.788(+/- 0.06)

Applying PCA

Random Forest Classification classification report:

	precision	recall	f1-score
Churn_No	0.93	0.93	0.93
Churn_Yes	0.93	0.93	0.93
avg / total	0.93	0.93	0.93

Cross Validation Accuracy Scores: 0.797(+/- 0.22)

Cross Validation Accuracy Scores - Test Set: 0.701(+/- 0.08)

Applying PCA

Decision Tree Classification classification report:

	precision	recall	f1-score
Churn_No	0.89	0.92	0.90
Churn_Yes	0.92	0.89	0.90
avg / total	0.90	0.90	0.90

Cross Validation Accuracy Scores: 0.804(+/- 0.23)

Cross Validation Accuracy Scores - Test Set: 0.728(+/- 0.04)

Applying PCA

K-Nearest Neighbours Classification classification report:

	precision	recall	f1-score
Churn_No	0.85	0.90	0.88
Churn_Yes	0.89	0.85	0.87
avg / total	0.87	0.87	0.87

Cross Validation Accuracy Scores: 0.815(+/- 0.20)

Cross Validation Accuracy Scores - Test Set: 0.737(+/- 0.06)

Applying PCA

Ridge Classification classification report:

	precision	recall	f1-score
Churn_No	0.64	1.00	0.78
Churn_Yes	1.00	0.45	0.62
avg / total	0.82	0.72	0.70

Cross Validation Accuracy Scores: 0.726(+/- 0.32)

Cross Validation Accuracy Scores - Test Set: 0.788(+/- 0.07)

Applying PCA

Lasso Classification classification report:

	precision	recall	f1-score
Churn_No	0.78	0.90	0.83
Churn_Yes	0.88	0.74	0.81
avg / total	0.83	0.82	0.71

Cross Validation Accuracy Scores: 0.809(+/- 0.27)

Cross Validation Accuracy Scores - Test Set: 0.789(+/- 0.05)

Applying PCA

Support Vector Classification (SVC) classification report:

	precision	recall	f1-score
Churn_No	0.80	0.89	0.84
Churn_Yes	0.88	0.77	0.82
avg / total	0.84	0.83	0.83

Cross Validation Accuracy Scores: 0.773(+/- 0.34)

Cross Validation Accuracy Scores - Test Set: 0.738(+/- 0.04)

Applying PCA

Gradient Boost Classification classification report:

	precision	recall	f1-score
Churn_No	0.93	0.94	0.93
Churn_Yes	0.94	0.92	0.93
avg / total	0.93	0.93	0.93

Cross Validation Accuracy Scores: 0.823(+/- 0.22)

Cross Validation Accuracy Scores - Test Set: 0.751(+/- 0.06)

Predicting churn using SelectKBest. >>

best k-features

SelectKBest removes all but the k highest scoring features

Applying SelectKBest

Logistic Regression (Classification L2) classification report:

	precision	recall	f1-score
Churn_No	0.83	0.88	0.85
Churn_Yes	0.88	0.82	0.84
avg / total	0.85	0.85	0.85

Cross Validation Accuracy Scores: 0.607(+/- 0.02)

Cross Validation Accuracy Scores - Test Set: 0.783(+/- 0.04)

Best score 0.7971538832609745
Best k-value {'kbest__k': 47}

Applying SelectKBest

Random Forest Classification classification report:

	precision	recall	f1-score
Churn_No	0.86	0.90	0.88
Churn_Yes	0.90	0.85	0.87
avg / total	0.88	0.88	0.88

Cross Validation Accuracy Scores: 0.739(+/- 0.10)

Cross Validation Accuracy Scores - Test Set: 0.764(+/- 0.04)

Best score 0.7842498794018331
Best k-value {'kbest__k': 39}

Applying SelectKBest

Decision Tree Classification classification report:

	precision	recall	f1-score
Churn_No	0.86	0.88	0.87
Churn_Yes	0.88	0.85	0.87
avg / total	0.87	0.87	0.87

Cross Validation Accuracy Scores: 0.740(+/- 0.10)

Cross Validation Accuracy Scores - Test Set: 0.763(+/- 0.04)

Best score 0.7957067052580801
Best k-value {'kbest__k': 45}

Applying SelectKBest

K-Nearest Neighbours Classification classification report:

	precision	recall	f1-score
Churn_No	0.84	0.88	0.86
Churn_Yes	0.88	0.83	0.85
avg / total	0.86	0.86	0.86

Cross Validation Accuracy Scores: 0.708(+/- 0.16)

Cross Validation Accuracy Scores - Test Set: 0.758(+/- 0.05)

Best score 0.787867824409069
Best k-value {'kbest__k': 30}

Applying SelectKBest

Ridge Classification classification report:

	precision	recall	f1-score
Churn_No	0.66	1.00	0.80
Churn_Yes	1.00	0.49	0.66
avg / total	0.83	0.75	0.73

Cross Validation Accuracy Scores: 0.606(+/- 0.02)

Cross Validation Accuracy Scores - Test Set: 0.783(+/- 0.05)

Best score 0.7589242643511819
Best k-value {'kbest__k': 46}

Applying SelectKBest

Lasso Classification classification report:

	precision	recall	f1-score
Churn_No	0.83	0.88	0.86
Churn_Yes	0.87	0.82	0.85
avg / total	0.85	0.85	0.85

Cross Validation Accuracy Scores: 0.607(+/- 0.02)

Cross Validation Accuracy Scores - Test Set: 0.782(+/- 0.04)

Best score 0.791365171249397
Best k-value {'kbest__k': 48}

Applying SelectKBest

Support Vector Classification (SVC) classification report:

	precision	recall	f1-score
Churn_No	0.82	0.88	0.86
Churn_Yes	0.88	0.80	0.84
avg / total	0.85	0.85	0.85

Cross Validation Accuracy Scores: 0.726(+/- 0.08)

Cross Validation Accuracy Scores - Test Set: 0.791(+/- 0.05)

Best score: 0.7969126869271587
Best k-value: {'kbest__k': 34}

Applying SelectKBest

Gradient Boost Classification classification report:

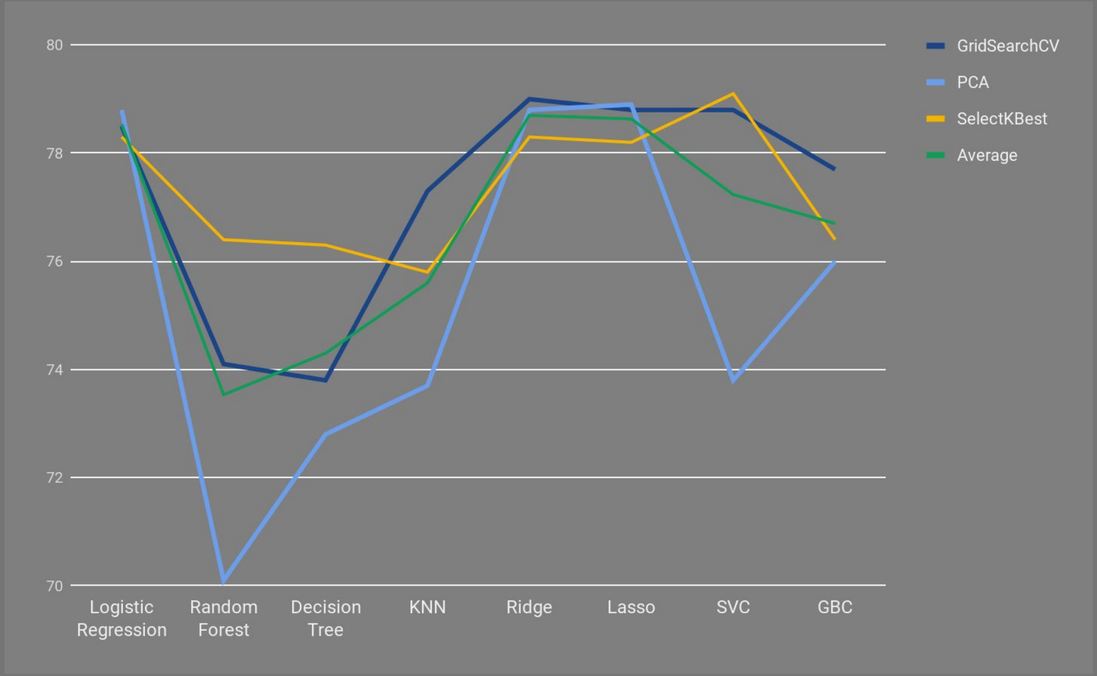
	precision	recall	f1-score
Churn_No	0.84	0.89	0.87
Churn_Yes	0.89	0.83	0.86
avg / total	0.86	0.86	0.86

Cross Validation Accuracy Scores: 0.607(+/- 0.02)

Cross Validation Accuracy Scores - Test Set: 0.782(+/- 0.04)

Best score: 0.8001688337433671
Best k-value: {'kbest__k': 50}

Overall Performance



Overall Performance

	GridSearchCV	PCA	SelectKBest
Logistic Regression	78.5	78.8	78.3
Random Forest	74.1	70.1	76.4
Decision Tree	73.8	72.8	76.3
KNN	77.3	73.7	75.8
Ridge C.	79	78.8	78.3
Lasso C.	78.8	78.9	78.2
SVC	78.8	73.8	79.1
Gradient Boost C.	77.0	76.0	76.4

Best model performance for predicting customer churn.

Logistic Regression	78.53 %
Random Forest C.	73.53 %
Decision Tree	74.30 %
KNN	75.60 %
Ridge C.	78.70 %
Lasso C.	78.63 %
Support Vector C.	77.23 %
Gradient Boost C.	76.70 %

Most important features and variables in customer churn.

>>

	Coefficient
Phone Service - Yes	2.172895
Monthly Charges - Medium	0.6554275
Multiple Lines - No phone service	0.062850

Probability of customers leaving.

Probability	Customer count	Predicted probability (mean)	True probability (mean)
0 - 10 %	508	0.038181	0.047074
10 - 20 %	187	0.145438	0.130854
20 - 30 %	156	0.250726	0.243905
30 - 40 %	118	0.349304	0.312120
40 - 50 %	119	0.448223	0.411362
50 - 60 %	109	0.547972	0.521068
60 - 70 %	81	0.649067	0.638258
70 - 80 %	80	0.749983	0.725620
80 - 90 %	48	0.832671	0.825686
90 - 100 %	1	0.908898	1.00000

How can this information help us?

Decide into which clients we should focus our resources

363 clients have over 70% probability of leaving

Address the most important issues that influence churn

Phone service, monthly charges, multiple lines

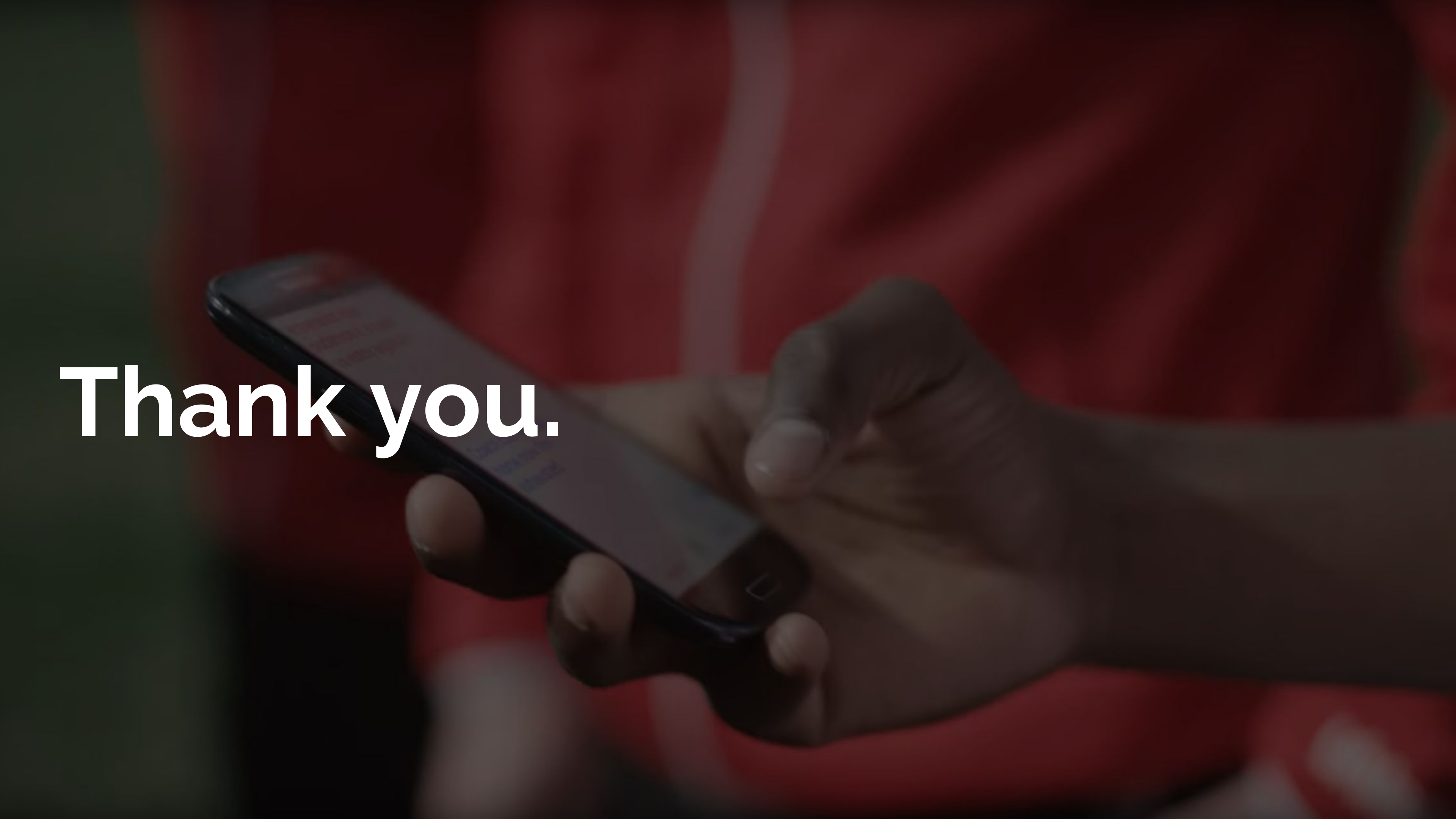
Reduce significantly campaign costs

Budgets are tight and we need to maximize our resources

Aspects we could improve to increase our predictability.

>>

- Dataset size
- Hyper parameter tuning
- Multiple algo testing
- Ensemble models
- Surveying for more features



Thank you.