Springboard

Customer Churn Prediction in Telecom

Capstone Project 1

Alex Martkovich

Executive Summary

A telecom company is looking to improve customer retention and decided develop focused customer retention programs based on customer churn prediction model.

The project team developed and tested several models, including:

- Linear regression
- Logistics regression
- Naïve bayes
- Random forest
- Random forest (upsampled)
- Random forest (downsampled)
- Stacked model

Random forest model tuning has been tested in order to improve prediction accuracy.

Developed predictive models may help **reduce** potential **revenue loss** due to churning customers **by 53%** while retaining up to 87% of customers.

Agenda

- **Introduction**
- **Historical data review**
- **Confusion matrix**
- **Base models**
- **Model tuning**
- **Up– and down– sampling**
- **Model stacking**
- **Business Case**

Introduction

A telecom company is looking to improve customer retention and decided develop focused customer retention programs based on customer churn prediction model.

The historical data includes information about:

- Customers who left within the last month
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been the company, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

In this document we've developed a few models, applied several techniques to improve prediction accuracy and built a business case.

NOTE: The analysis did not take into account probability of the customer retention after it has been correctly predicted.

Agenda

- Introduction
- Historical data review
- Confusion matrix
- Base models
- Model tuning
- Up– and down– sampling
- Model stacking
- Business Case

The Data contains 19 features and a target

#	Field	Description	Example
1	customerID	Customer ID	7590-VHVEG
2	gender	Whether the customer is a male or a female	Female
3	SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)	0
4	Partner	Whether the customer has a partner or not (Yes, No)	Yes
5	Dependents	Whether the customer has dependents or not (Yes, No)	No
6	tenure	Number of months the customer has stayed with the company	1
7	PhoneService	Whether the customer has a phone service or not (Yes, No)	No
8	MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)	No phone service
9	InternetService	Customer's internet service provider (DSL, Fiber optic, No)	DSL
10	OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)	No
11	OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)	Yes
12	DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)	No
13	TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)	No
14	StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)	No
15	StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)	No
16	Contract	The contract term of the customer (Month-to-month, One year, Two year)	Month-to-month
17	PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)	Yes
18	PaymentMethod	Payment method - Electronic check, Mailed check, Bank transfer, Credit card	Electronic check
19	MonthlyCharges	The amount charged to the customer monthly	29.85
20	TotalCharges	The total amount charged to the customer	29.85
21	Churn (Target)	Whether the customer churned or not (Yes or No)	No

Dataset size and preprocessing

- The dataset has no missing values and requires no data cleaning

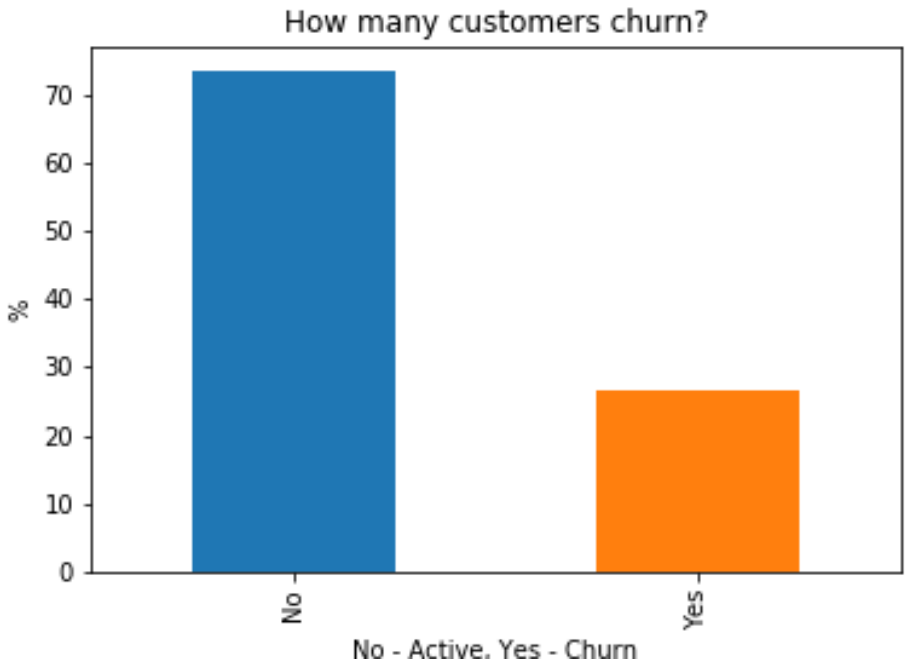
- Overall the file contains 7043 records and 20 columns

- In order to be able to perform modeling, the data requires preprocessing, including:
  - Converting string format to numerical
  - Converting categorical data to dummy values
  - Converting binary categorical to numerical

- After preprocessing the number of feature fields increased from 19 to 40
- Considering that the file contains significant amount of features, but no time series data or behavioral data, the feature engineering has not been done

Target variable

Approximately 26% of the customers churn



- Due to underrepresented data on Churning customers we can assume that the model will perform better at predicting non-churning customers

- We will test strategies to deal with this imbalance

Agenda

- Introduction
- Historical data review
- Confusion matrix
- Base models
- Model tuning
- Up– and down– sampling
- Model stacking
- Business Case

Confusion matrix – we are looking to improve precision and recall of churn prediction

		Predicted value	
		1 (Positive)	**0 (Negative)**
Actual value	**1 Positive**	**TP** **Customers who are predicted and will churn**	FN Customers who are predicted to stay, but will churn
	0 Negative	**FP** **Customers who are predicted to churn, but will stay**	TN Customers who are predicted and will stay

Analysis Focus

Net effect of predicted model parameters

Importance	Value	Event	Treatment cost	Retained Profit	Net effect	Objective
1	**TP**	Customers who are predicted and will churn	-$422*	$2,279**	$1,875	Maximize
2	**FP**	Customers who are predicted to churn, but will stay	-$422	0	-$422	Minimize
3	FN	Customers who are predicted to stay, but will churn	0	0	0	
4	TN	Customers who are predicted and will stay	0	0	0	

Analysis Focus

* Average monthly revenue per customer ~ $70.42. Offered promotion to retain a customer is 50% discount for 6 months ~ $422.55

** Average revenue per customer to date ~ $2,279

Model performance assessment

The model performance have been compared and optimized based on the following parameters:

1. **Recall** – number of accurately predicted churning customers out of total number of churning customers, to offer treatment and reduce forgone revenue (the higher Recall the better)

2. **Precision** – number of customers predicted to churn, but will actually stay, to reduce the total number and cost of offered treatment (the higher Precision the better)

3. **Revenue retained** – amount of revenue retained due to churn prediction model application out of the total potential forgone revenue

4. **Accuracy score** – overall model accuracy

Agenda

- Introduction
- Historical data review
- Confusion matrix
- Base models
- Model tuning
- Up– and down– sampling
- Model stacking
- Business Case

4 base models have been tested

The following base models have been tested:

- Linear regression
- Logistics regression
- Naïve Bayes
- Random Forest

Linear regression

Confusion matrix

		Predicted value	
		1 Churn	0
Actual value	1 Churn	309	265
	0	136	1403

- Recall (1) : 0.54
- Precision (1): 0.69
- Accuracy score: 0.81
- **Revenue Retained: $516,403**
- **Revenue Retained: 39%**

Logistics regression

Confusion matrix

		Predicted value	
		1 Churn	**0**
Actual value	**1 Churn**	**330**	**244**
	0	158	1381

- Recall (1) : 0.57
- Precision (1): 0.68
- Accuracy score: 0.81
- **Revenue Retained: $546,108**
- **Revenue Retained: 42%**

Naive Bayes

Confusion matrix

		Predicted value	
		1 Churn	**0**
Actual value	**1 Churn**	**500**	**74**
	0	561	978

- Recall (1) : 0.87
- Precision (1): 0.47
- Accuracy score: 0.70
- **Revenue Retained: $691,541**
- **Revenue Retained: 53%**

Random Forest

Confusion matrix

		Predicted value	
		1 Churn	0
Actual value	1 Churn	279	295
	0	137	1402

- Recall (1) : 0.49
- Precision (1): 0.67
- Accuracy score: 0.80
- **Revenue Retained: $460,265**
- **Revenue Retained: 35%**

Base model comparison – Naïve Bayes model produces the best financial impact

Model	Recall	Precision	Accuracy	Revenue Retained	Revenue Retained %
Linear Regression	0.54	0.69	0.81	$516,403	39%
Logistics Regression	0.57	0.68	0.81	$546,108	42%
Naïve Bayes	0.87	0.47	0.70	$691,542	53%
Random Forest	0.49	0.67	0.80	$460,265	35%

Agenda

- Introduction
- Historical data review
- Confusion matrix
- Base models
- Model tuning
- Up– and down– sampling
- Model stacking
- Business Case

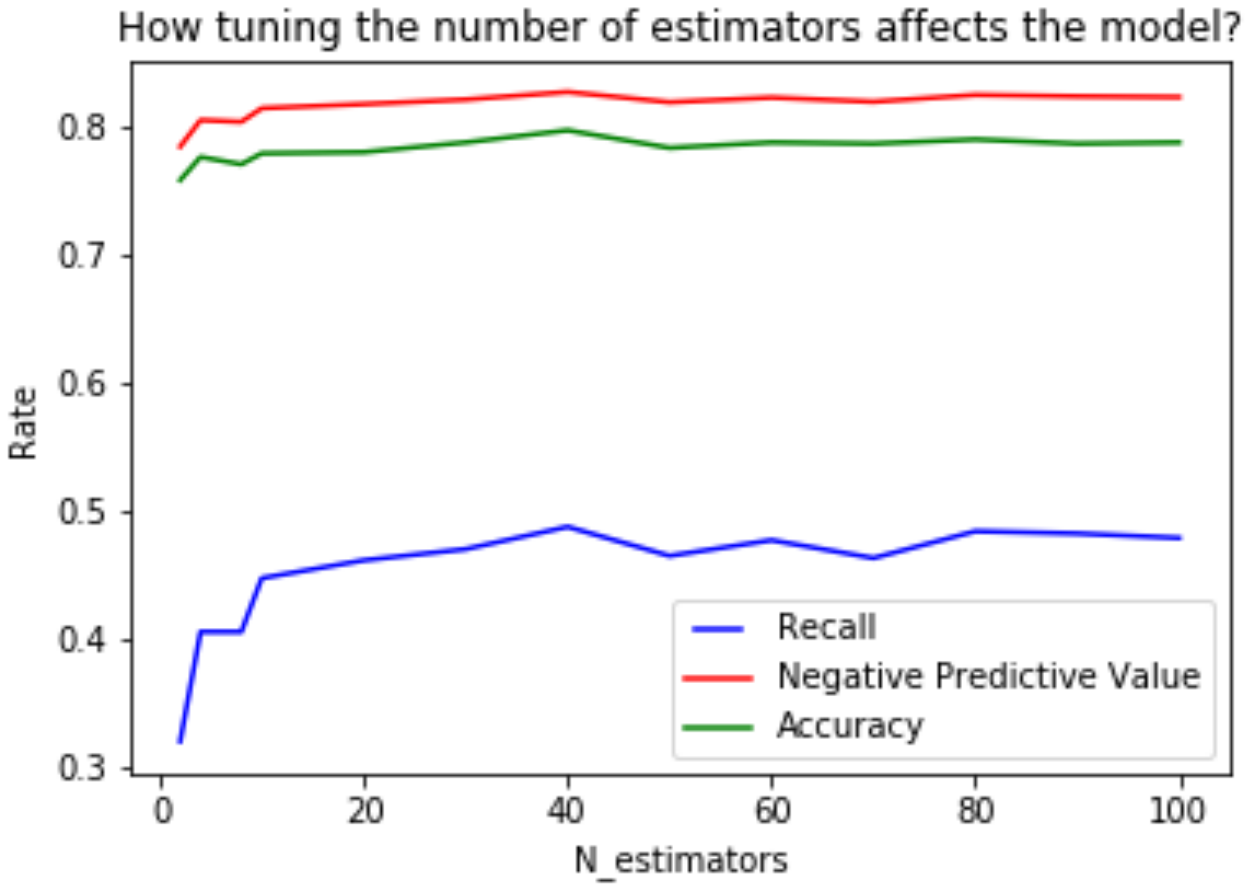
Random Forest tuning

We performed tuning of the Random Forest model parameters to evaluate how they impact the model accuracy

We have tuned the following parameters:

- n_estimators - the number of trees in the forest
  Usually the higher the number of trees the better. However, adding more trees can slow down the training process
- min_samples_leaf - the minimum number of samples required to be at a leaf node
  This parameter describes the minimum number of samples at the leafs, the base of the tree
- max_depth - the depth of each tree in the forest
  The deeper the tree, the more splits it has and it captures more information about the data
- max_features - the number of features to consider when looking for the best split

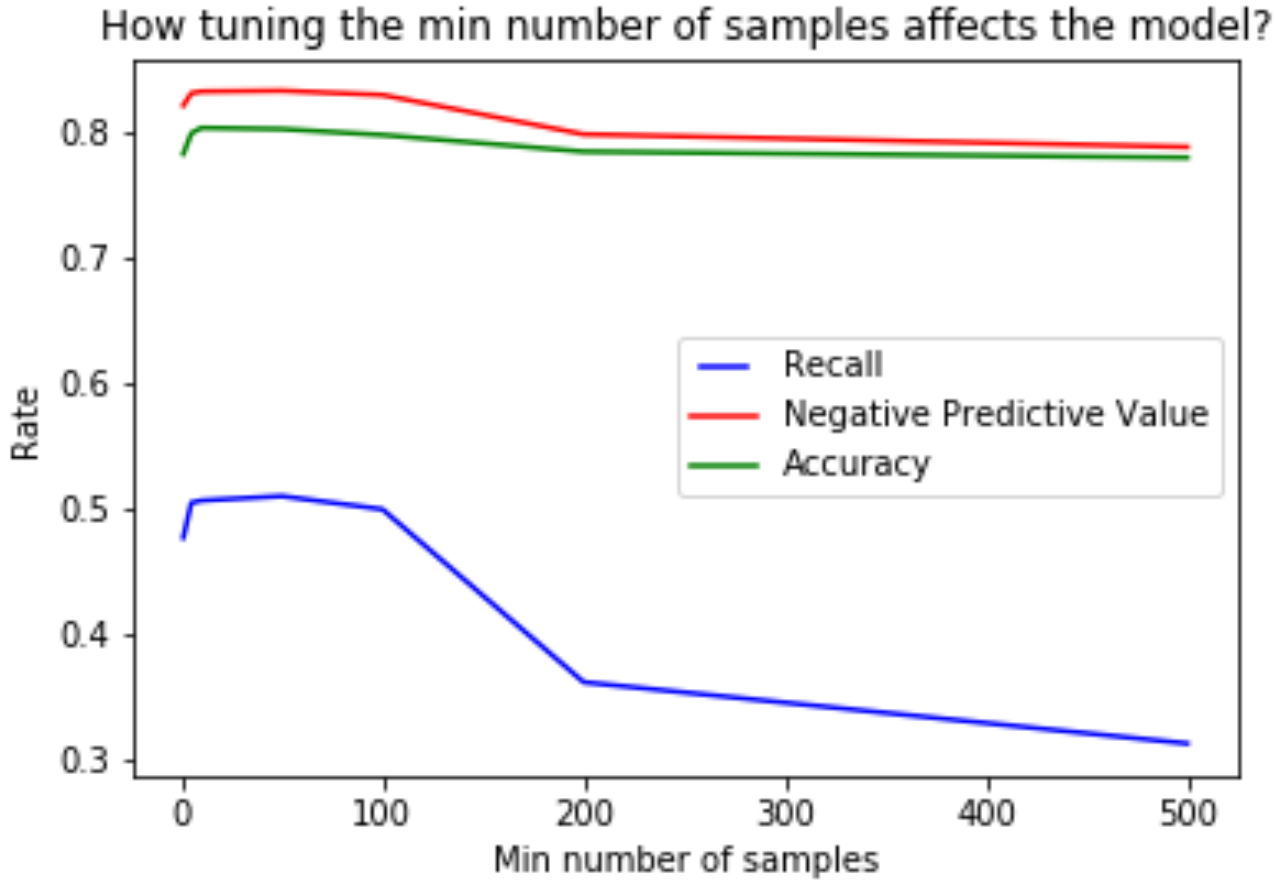
Number of trees in the forest



Comments

- We can see that for this dataset, we can stop at 40 trees as increasing further the number of trees decreases the test performance

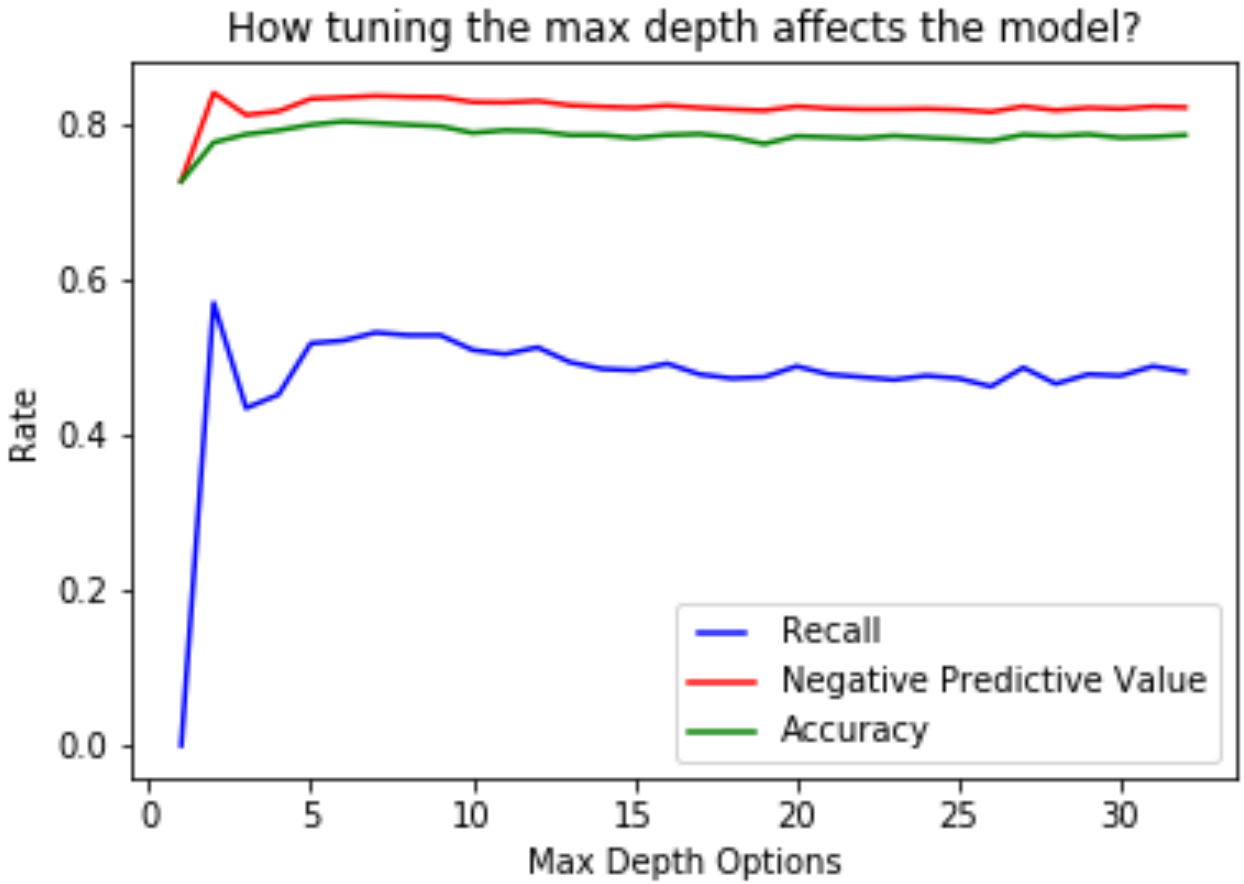
Minimum number of samples required to split an internal node



Comments

- We can see that when we increase the number samples above 100, the model cannot learn enough about the data
- Increasing this value can cause underfitting

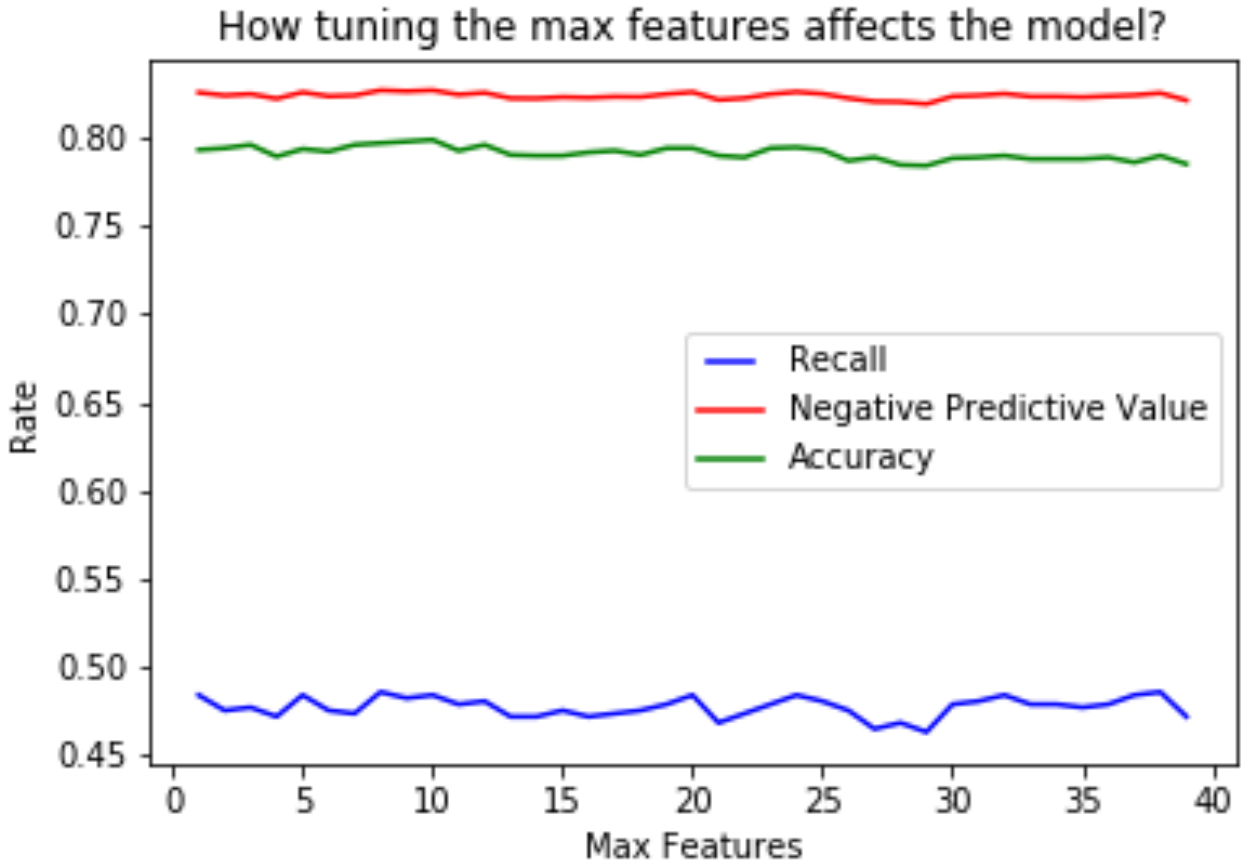
Depth of each tree in the forest



Comments

- We see that the model overfits for larger depth values
- The tree perfectly predicts all of the train data, however, it fails to generalize the findings for new data

Number of features to consider when looking for the best split



Comments

- We can see that the model is not significantly affected by tuning the max number of features

Agenda

- Introduction
- Historical data review
- Confusion matrix
- Base models
- Model tuning
- Up- and down-sampling
- Model stacking
- Business Case

Dealing with imbalanced data

- Our models predict a two-class problem, that has a binary response
- As mentioned earlier, 74% of the available training data represents the **majority class** of the target variable – non-churn
- Whereas we are trying to predict customer churn, which is a **minority class** in this dataset
- One of the most common ways to address the imbalance is to create a balanced training set with equal numbers of minority and majority class examples
- We will test two ways of doing this:
  - Up-sampling, and
  - Down-sampling
- We will apply this technique to the Random Forest model

Random Forest - upsampling

Confusion matrix

		Predicted value	
		1 Churn	0
Actual value	1 Churn	**359**	**215**
	0	268	1271

Results (upsampled vs base model)**:**

- Recall (1) : 0.63 vs 0.49
- Precision (1): 0.57 vs 0.67
- Accuracy score: 0.77 vs 0.80
- **Revenue Retained: $553,486** vs $460,265
- **Revenue Retained: 42%**

Comments:

- The following approach has been used:
  - Upsampled the minority class from 1295 up to 3635 records
  - Trained the model on the upsampled data
  - Applied the trained model to the original test dataset
- Overall accuracy is lower than the base model due to higher number of FP
- But the model produces **better positive financial impact** as compared to the base model, due to better prediction of TP

Random Forest - downsampling

Confusion matrix

		Predicted value	
		1 Churn	0
Actual value	1 Churn	**459**	**115**
	0	440	1099

Results (downsampled vs base model)**:**

- Recall (1) : 0.80 vs 0.49
- Precision (1): 0.51 vs 0.67
- Accuracy score: 0.74 vs 0.80
- **Revenue Retained: $666,526** vs $460,265
- **Revenue Retained: 51%**

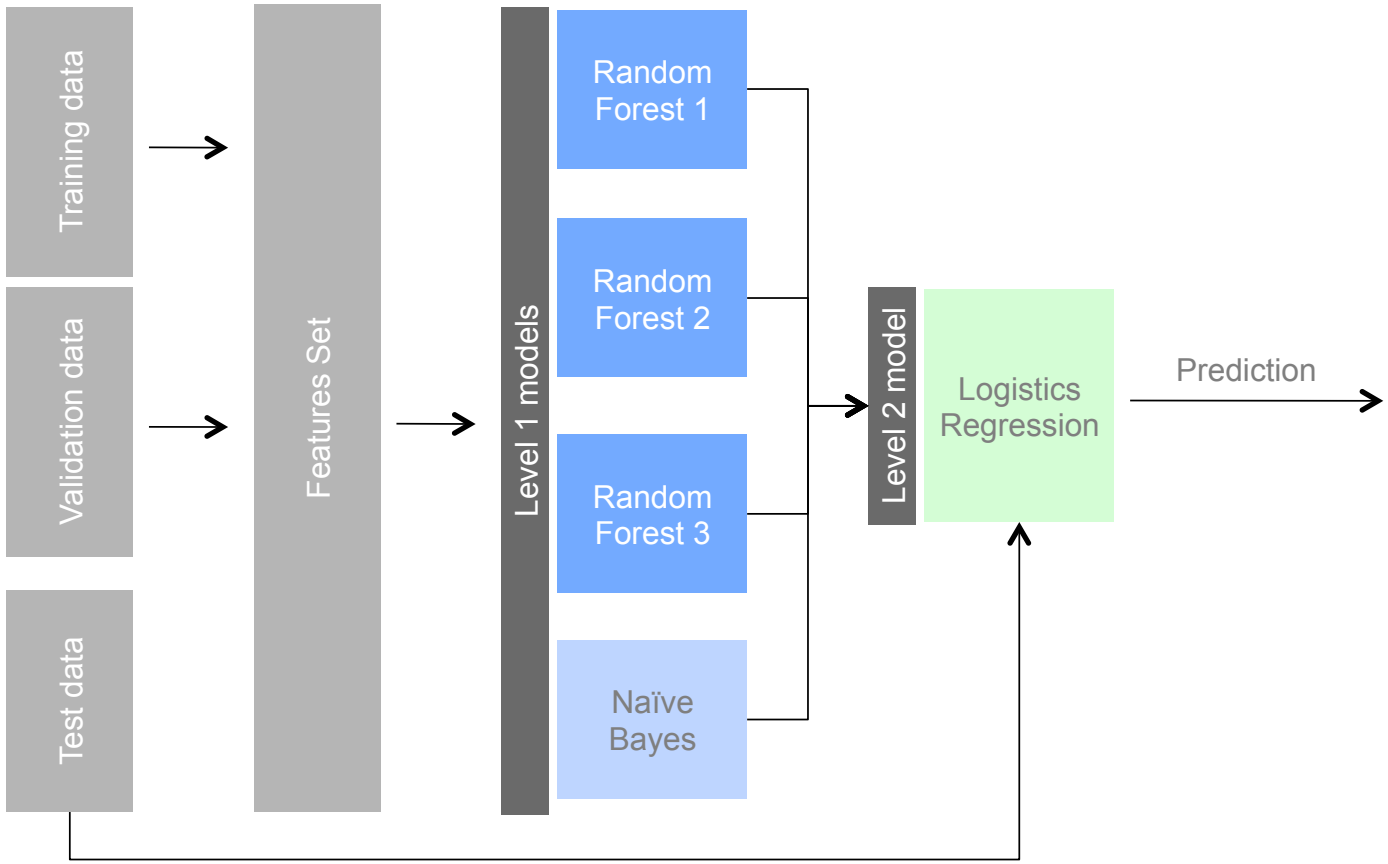
Comments:

- The following approach has been used:
  - Downsampled the majority class from 3635 down to 1295 records
  - Trained the model on the downsampled data
  - Applied the trained model to the original test dataset
- The model accuracy is lower as compared to the base and upsampled model due to higher number of False Positive values
- The model produces **better financial impact** as compared to the upsampled and base models due to better prediction of FP

Agenda

- Introduction
- Historical data review
- Confusion matrix
- Base models
- Model tuning
- Up– and down– sampling
- Model stacking
- Business Case

Model stacking was performed on 2-level model



Comments

- 3 Random Forest models with different settings and a Naïve Bayes model were used in the 1st level
- Logistics Regression have been used for the 2nd level

Model stacking results

Confusion matrix

		Predicted value	
		1 Churn	0
Actual value	1 Churn	241	232
	0	176	1,363

- Recall (1) : 0.60
- Precision (1): 0.66
- Accuracy score: 0.81
- **Revenue Retained: $560,788**
- **Revenue Retained: 43%**

Agenda

- Introduction
- Historical data review
- Confusion matrix
- Base models
- Model tuning
- Up– and down– sampling
- Model stacking
- Business Case

Customer churn prediction may save up to 53% of potentially forgone revenue

Model	Recall	Precision	Accuracy	Revenue Retained	Revenue Retained %
Naïve Bayes (Option 1)	0.87	0.47	0.70	$691,542	53%
Random Forest (Option 2) Downsampled	0.80	0.51	0.74	$666,528	51%
Stacked Model	0.60	0.66	0.81	$560,788	43%
Random Forest Upsampled	0.63	0.57	0.77	$553,486	42%
Logistics Regression	0.57	0.68	0.81	$546,108	42%
Linear Regression	0.54	0.69	0.81	$516,403	39%
Random Forest	0.49	0.67	0.80	$460,265	35%

Predictive models may reduce revenue loss by 53% while retaining up to 87% of customers

Recommended options	As Is	Option 2: Random Forest (Downsampled)	Option 1: **Naïve Bayes**
Customers churned	574	115	**74**
Customers retained	0	459	**500**
Revenue loss due to customer churn	-$1,308,146	-$642,041	**-$617,025**
Revenue retained	$0	666,526	**$691,541**
Customers retained, %	0%	80%	**87%**
Revenue retained, %	0%	51%	**53%**

Recommendation:

- Based on the results of the business case, it is advised to develop and apply the customer churn prediction models to reduce revenue loss due to customer churn.
- Naïve Bayes model provided the best prediction results and financial impact.
- NOTE: The analysis does not take into account probability of the customer retention after it has been correctly predicted and the actual results could be different


Springboard