

HarvardX - Data Science Capstone: Win/Loss Analysis Project

1. Introduction

Let's consider a real life scenario where we play the role of a sales executive at an automotive supply wholesaler and investigate a sales execution issue.

We have not been converting enough opportunities lately. We want to better **understand our sales pipeline and which deals our sales teams can expect to win or lose** based on data that we've pulled out of our CRM database.

We want to find the patterns in sales wins and losses and uncover what can lead to successful sales opportunities and better anticipate performance gaps.

2. Data

The dataset is a sample provided by IBM in their Watson Analytics community that can be downloaded [here](#). The "WA_Fn UseC_Sales Win Loss.csv" file is a dataset that covers sales activities for, amongst others, carrying out a win/loss analysis, to generate the **insights to increase revenues and grow the business**.

Dataset Features

Column name	Description
Client Size by Employee Count	Employee sized by number of clients: • 1: < 1K • 2: [1K, 5K] • 3: [5K, 10K] • 4: [10K, 30K] • 5: 30K
Client Size by Revenue	Client size based on annual revenue in USD: • 1: < 1M • 2: [1M, 10M] • 3: [10M, 50M] • 4: [50M, 100M] • 5: 100M
Competitor Type	An indicator if a competitor has been identified: Known, Unknown, None
Deal Size by Category	Categorical grouping of the opportunity amount (OpportunityAmountUSD) • 1: < 10K • 2: [10K, 25K] • 3: [25K, 50K] • 4: [50K, 100K] • 5: [100K, 250K] • 6: [250K, 500K] • 7: 500K
Opportunity Number	A unique generated number assigned to the opportunity
Opportunity Results	A closed opportunity is won or loss. Values could be Win/Loss
Region	Name of the Region: Mid-Atlantic, Midwest, Northeast, Northwest, Pacific, Southeast, Southwest
Route to Market	The opportunities' route to market: Fields Sales, Other, Reseller, Telecoverage, Telesales
Supplies Group	Reporting supplies group: Car Accessories, Car Electronics, Performance & Non-auto, Tires & Wheels
Supplies SubGroup	Reporting supplies subgroup: Batteries & Accessories, Car Electronics, Exterior Accessories, Garage & Car Care, Interior Accessories, Motorcycle Parts, Performance Parts, Replacement Parts, Shelters & RV, Tires & Wheels, Towing & Hitches
Opportunity Amount (USD)	Sum of line item revenue estimates by sales representative in American currency

Column name	Description
Sales Stage Change Count	Actually a count of number of times an opportunity changes sales stages (back and forwards)
Elapsed Days In Sales Stage	The number of days between the change in sales stages. The counter is reset for each new sales stage
Ratio Days Identified To Total Days	Ratio of total days the opportunity has spent in sales stage: Identified/Validating over total days in sales process
Ratio Days Qualified To Total Days	Ratio of total days the opportunity has been spent in sales stage: Qualified/Gaining Agreement over total days in sales process
Ratio Days Validated To Total Days	Ratio of total days the Opportunity has presence in sales stage: Validated/Qualifying over total days in sales process
Revenue From Client Past Two Years	Revenue identified from this client in past two years • 0: 0 • 1: [1K, 50K] • 2: [50K, 400K] • 3: [400K, 1.5M] • 4: 1.5M
Total Days Identified Through Closing	Total days the opportunity has spent in Sales Stages from Identified/Validating to Gained Agreement/closing
Total Days Identified Through Qualified	Total days the opportunity has spent in CRM Stages from Identified/Validating to Qualified/Gaining Agreement

3. Methodology

3.1 Exploratory Data Analysis

In this section, we explore the data in two main steps:

- **Initial exploration**
 - Dataset structure, variable formats,
 - Missing values,
 - Duplicated information,
 - Correlation analysis.
- **In-depth exploration for first insights**

3.2 Create subsets for the project

We want to create two subsets as follows:

- sales dataset, which contains 90% of our sample dataset, to analyze our sales wins and losses.
- validation dataset, which is the remaining 10%, for the purpose of validation of our predictive model.

3.3 Predictive Model

In this section, we will go through a couple of Machine Learning methods to build a model to support our decisions.

4. Results and Discussion

4.1 Exploratory Data Analysis

Initial data exploration

##	Opportunity.Number	Supplies.Subgroup	Supplies.Group	Region
## 1	1641984	Exterior Accessories	Car Accessories	Northwest
## 2	1658010	Exterior Accessories	Car Accessories	Pacific
## 3	1674737	Motorcycle Parts	Performance & Non-auto	Pacific
## 4	1675224	Shelters & RV	Performance & Non-auto	Midwest
## 5	1689785	Exterior Accessories	Car Accessories	Pacific
## 6	1692390	Shelters & RV	Performance & Non-auto	Pacific
##	Route.To.Market	Elapsed.Days.In.Sales.Stage	Opportunity.Result	
## 1	Fields Sales	76	Won	
## 2	Reseller	63	Loss	
## 3	Reseller	24	Won	
## 4	Reseller	16	Loss	
## 5	Reseller	69	Loss	
## 6	Reseller	89	Loss	
##	Sales.Stage.Change.Count	Total.Days.Identified.Through.Closing		
## 1	13	104		
## 2	2	163		
## 3	7	82		
## 4	5	124		
## 5	11	91		
## 6	3	114		
##	Total.Days.Identified.Through.Qualified	Opportunity.Amount.USD		
## 1	101	0		
## 2	163	0		
## 3	82	7750		
## 4	124	0		
## 5	13	69756		
## 6	0	232522		
##	Client.Size.By.Revenue	Client.Size.By.Employee.Count		
## 1	5	5		
## 2	3	5		
## 3	1	1		
## 4	1	1		
## 5	1	1		
## 6	5	1		
##	Revenue.From.Client.Past.Two.Years	Competitor.Type		
## 1	0	Unknown		
## 2	0	Unknown		
## 3	0	Unknown		
## 4	0	Known		
## 5	0	Unknown		
## 6	0	Unknown		
##	Ratio.Days.Identified.To.Total.Days	Ratio.Days.Validated.To.Total.Days		
## 1	0.69636	0.113985		
## 2	0.00000	1.000000		
## 3	1.00000	0.000000		
## 4	1.00000	0.000000		
## 5	0.00000	0.141125		
## 6	0.00000	0.000877		
##	Ratio.Days.Qualified.To.Total.Days	Deal.Size.Category		
## 1	0.154215	1		
## 2	0.000000	1		
## 3	0.000000	1		
## 4	0.000000	1		

```

## 5          0.000000          4
## 6          0.000000          5

## 'data.frame':  78025 obs. of  19 variables:
## $ Opportunity.Number      : int  1641984 1658010 1674737 1675224 1689785 1692390 193
## $ Supplies.Subgroup       : Factor w/ 11 levels "Batteries & Accessories",...: 3 3 6
## $ Supplies.Group          : Factor w/ 4 levels "Car Accessories",...: 1 1 3 3 1 3 1
## $ Region                   : Factor w/ 7 levels "Mid-Atlantic",...: 4 5 5 2 5 5 5 4
## $ Route.To.Market          : Factor w/ 5 levels "Fields Sales",...: 1 3 3 3 3 3 1 1 1
## $ Elapsed.Days.In.Sales.Stage : int  76 63 24 16 69 89 111 82 68 18 ...
## $ Opportunity.Result       : Factor w/ 2 levels "Loss","Won": 2 1 2 1 1 1 2 1 1 1 ...
## $ Sales.Stage.Change.Count  : int  13 2 7 5 11 3 12 6 8 7 ...
## $ Total.Days.Identified.Through.Closing : int  104 163 82 124 91 114 112 70 156 50 ...
## $ Total.Days.Identified.Through.Qualified: int  101 163 82 124 13 0 112 70 156 50 ...
## $ Opportunity.Amount.USD    : int  0 0 7750 0 69756 232522 20001 450000 250000 55003 ...
## $ Client.Size.By.Revenue    : int  5 3 1 1 1 5 4 1 1 1 ...
## $ Client.Size.By.Employee.Count : int  5 5 1 1 1 1 5 1 5 1 ...
## $ Revenue.From.Client.Past.Two.Years : int  0 0 0 0 0 0 0 0 0 0 ...
## $ Competitor.Type           : Factor w/ 3 levels "Known","None",...: 3 3 3 1 3 3 3 1 2
## $ Ratio.Days.Identified.To.Total.Days : num  0.696 0 1 1 0 ...
## $ Ratio.Days.Validated.To.Total.Days : num  0.114 1 0 0 0.141 ...
## $ Ratio.Days.Qualified.To.Total.Days : num  0.154 0 0 0 0 ...
## $ Deal.Size.Category        : int  1 1 1 1 4 5 2 6 6 4 ...

## 'data.frame':  78025 obs. of  19 variables:
## $ Opportunity.Number      : int  1641984 1658010 1674737 1675224 1689785 1692390 193
## $ Supplies.Subgroup       : Factor w/ 11 levels "Batteries & Accessories",...: 3 3 6
## $ Supplies.Group          : Factor w/ 4 levels "Car Accessories",...: 1 1 3 3 1 3 1
## $ Region                   : Factor w/ 7 levels "Mid-Atlantic",...: 4 5 5 2 5 5 5 4
## $ Route.To.Market          : Factor w/ 5 levels "Fields Sales",...: 1 3 3 3 3 3 1 1 1
## $ Elapsed.Days.In.Sales.Stage : int  76 63 24 16 69 89 111 82 68 18 ...
## $ Opportunity.Result       : Factor w/ 2 levels "Loss","Won": 2 1 2 1 1 1 2 1 1 1 ...
## $ Sales.Stage.Change.Count  : int  13 2 7 5 11 3 12 6 8 7 ...
## $ Total.Days.Identified.Through.Closing : int  104 163 82 124 91 114 112 70 156 50 ...
## $ Total.Days.Identified.Through.Qualified: int  101 163 82 124 13 0 112 70 156 50 ...
## $ Opportunity.Amount.USD    : int  0 0 7750 0 69756 232522 20001 450000 250000 55003 ...
## $ Client.Size.By.Revenue    : Factor w/ 5 levels "1","2","3","4",...: 5 3 1 1 1 5 4 1 1
## $ Client.Size.By.Employee.Count : Factor w/ 5 levels "1","2","3","4",...: 5 5 1 1 1 1 5 1 5
## $ Revenue.From.Client.Past.Two.Years : Factor w/ 5 levels "0","1","2","3",...: 1 1 1 1 1 1 1 1 1
## $ Competitor.Type           : Factor w/ 3 levels "Known","None",...: 3 3 3 1 3 3 3 1 2
## $ Ratio.Days.Identified.To.Total.Days : num  0.696 0 1 1 0 ...
## $ Ratio.Days.Validated.To.Total.Days : num  0.114 1 0 0 0.141 ...
## $ Ratio.Days.Qualified.To.Total.Days : num  0.154 0 0 0 0 ...
## $ Deal.Size.Category        : Factor w/ 7 levels "1","2","3","4",...: 1 1 1 1 4 5 2 6 6

## Do we have any missing value? FALSE

## We have 77829 unique opportunity numbers out of a total of 78025 so the percentage of duplicated rows is 0.0031

##      Opportunity.Number      Supplies.Subgroup      Supplies.Group
## 93      4947042      Exterior Accessories      Car Accessories
## 94      4947042      Towing & Hitches      Car Accessories
## 453     5629727      Shelters & RV Performance & Non-auto
## 454     5629727      Shelters & RV Performance & Non-auto
## 725     5799657      Interior Accessories      Car Accessories
## 726     5799657      Batteries & Accessories      Car Accessories

```

## 1105	5934206	Garage & Car Care	Car Accessories
## 1106	5934206	Batteries & Accessories	Car Accessories
## 1153	5943944	Batteries & Accessories	Car Accessories
## 1154	5943944	Batteries & Accessories	Car Accessories
##	Region	Route.To.Market	Elapsed.Days.In.Sales.Stage
## 93	Midwest	Fields Sales	88
## 94	Midwest	Fields Sales	88
## 453	Northwest	Fields Sales	89
## 454	Northwest	Fields Sales	89
## 725	Pacific	Fields Sales	83
## 726	Pacific	Fields Sales	83
## 1105	Pacific	Fields Sales	89
## 1106	Pacific	Fields Sales	89
## 1153	Midwest	Fields Sales	86
## 1154	Midwest	Fields Sales	86
##	Opportunity.Result	Sales.Stage.Change.Count	
## 93	Loss	4	
## 94	Loss	4	
## 453	Loss	2	
## 454	Loss	2	
## 725	Loss	3	
## 726	Loss	3	
## 1105	Loss	4	
## 1106	Loss	4	
## 1153	Loss	4	
## 1154	Loss	4	
##	Total.Days.Identified.Through.Closing		
## 93		81	
## 94		81	
## 453		51	
## 454		51	
## 725		49	
## 726		49	
## 1105		37	
## 1106		37	
## 1153		40	
## 1154		40	
##	Total.Days.Identified.Through.Qualified	Opportunity.Amount.USD	
## 93		81	600000
## 94		81	200000
## 453		51	200000
## 454		51	0
## 725		49	250000
## 726		49	350000
## 1105		37	25000
## 1106		37	26000
## 1153		40	30000
## 1154		40	5000
##	Client.Size.By.Revenue	Client.Size.By.Employee.Count	
## 93		3	5
## 94		3	5
## 453		1	5
## 454		1	5
## 725		4	1

## 726	4	1
## 1105	1	1
## 1106	1	1
## 1153	1	3
## 1154	1	3

##	Revenue.From.Client.Past.Two.Years	Competitor.Type
## 93	3	Known
## 94	3	Known
## 453	0	Unknown
## 454	0	Unknown
## 725	0	Unknown
## 726	0	Unknown
## 1105	0	Unknown
## 1106	0	Unknown
## 1153	0	Known
## 1154	0	Known

##	Ratio.Days.Identified.To.Total.Days
## 93	0.170972
## 94	0.170972
## 453	0.000000
## 454	0.000000
## 725	0.556911
## 726	0.556911
## 1105	0.597297
## 1106	0.597297
## 1153	0.020000
## 1154	0.020000

##	Ratio.Days.Validated.To.Total.Days	Ratio.Days.Qualified.To.Total.Days
## 93	0.030750	0.798278
## 94	0.030750	0.798278
## 453	1.000000	0.000000
## 454	1.000000	0.000000
## 725	0.443089	0.000000
## 726	0.443089	0.000000
## 1105	0.402703	0.000000
## 1106	0.402703	0.000000
## 1153	0.235000	0.745000
## 1154	0.235000	0.745000

##	Deal.Size.Category
## 93	7
## 94	5
## 453	5
## 454	1
## 725	6
## 726	6
## 1105	3
## 1106	3
## 1153	3
## 1154	1

Do we have any missing value? FALSE

We have 77829 unique opportunity numbers out of a total of 77829 so the percentage of duplicated rows

Opportunity.Number

## Opportunity.Number	1.000000000
## Elapsed.Days.In.Sales.Stage	-0.764971087
## Sales.Stage.Change.Count	-0.253592731
## Total.Days.Identified.Through.Closing	-0.445872151
## Total.Days.Identified.Through.Qualified	-0.435432828
## Opportunity.Amount.USD	-0.014999049
## Ratio.Days.Identified.To.Total.Days	0.002240132
## Ratio.Days.Validated.To.Total.Days	-0.057535839
## Ratio.Days.Qualified.To.Total.Days	-0.048404076
## Elapsed.Days.In.Sales.Stage	
## Opportunity.Number	-0.76497109
## Elapsed.Days.In.Sales.Stage	1.00000000
## Sales.Stage.Change.Count	-0.02533635
## Total.Days.Identified.Through.Closing	-0.02267760
## Total.Days.Identified.Through.Qualified	-0.02339319
## Opportunity.Amount.USD	-0.01392492
## Ratio.Days.Identified.To.Total.Days	-0.01826893
## Ratio.Days.Validated.To.Total.Days	0.01582840
## Ratio.Days.Qualified.To.Total.Days	0.02040566
## Sales.Stage.Change.Count	
## Opportunity.Number	-0.25359273
## Elapsed.Days.In.Sales.Stage	-0.02533635
## Sales.Stage.Change.Count	1.00000000
## Total.Days.Identified.Through.Closing	0.38037554
## Total.Days.Identified.Through.Qualified	0.35572017
## Opportunity.Amount.USD	0.05040764
## Ratio.Days.Identified.To.Total.Days	-0.06813596
## Ratio.Days.Validated.To.Total.Days	-0.01320735
## Ratio.Days.Qualified.To.Total.Days	0.28067095
## Total.Days.Identified.Through.Closing	
## Opportunity.Number	-0.44587215
## Elapsed.Days.In.Sales.Stage	-0.02267760
## Sales.Stage.Change.Count	0.38037554
## Total.Days.Identified.Through.Closing	1.00000000
## Total.Days.Identified.Through.Qualified	0.98365127
## Opportunity.Amount.USD	0.07722813
## Ratio.Days.Identified.To.Total.Days	0.09249282
## Ratio.Days.Validated.To.Total.Days	0.12477671
## Ratio.Days.Qualified.To.Total.Days	0.02652307
## Total.Days.Identified.Through.Qualified	
## Opportunity.Number	-0.43543283
## Elapsed.Days.In.Sales.Stage	-0.02339319
## Sales.Stage.Change.Count	0.35572017
## Total.Days.Identified.Through.Closing	0.98365127
## Total.Days.Identified.Through.Qualified	1.00000000
## Opportunity.Amount.USD	0.07811019
## Ratio.Days.Identified.To.Total.Days	0.10474292
## Ratio.Days.Validated.To.Total.Days	0.14574872
## Ratio.Days.Qualified.To.Total.Days	0.02808720
## Opportunity.Amount.USD	
## Opportunity.Number	-0.01499905
## Elapsed.Days.In.Sales.Stage	-0.01392492
## Sales.Stage.Change.Count	0.05040764
## Total.Days.Identified.Through.Closing	0.07722813

```

## Total.Days.Identified.Through.Qualified      0.07811019
## Opportunity.Amount.USD                      1.00000000
## Ratio.Days.Identified.To.Total.Days          0.06732225
## Ratio.Days.Validated.To.Total.Days          -0.05182146
## Ratio.Days.Qualified.To.Total.Days          0.04672572
## Ratio.Days.Identified.To.Total.Days
## Opportunity.Number                          0.002240132
## Elapsed.Days.In.Sales.Stage                 -0.018268925
## Sales.Stage.Change.Count                   -0.068135965
## Total.Days.Identified.Through.Closing        0.092492819
## Total.Days.Identified.Through.Qualified      0.104742923
## Opportunity.Amount.USD                      0.067322251
## Ratio.Days.Identified.To.Total.Days          1.000000000
## Ratio.Days.Validated.To.Total.Days          -0.476667451
## Ratio.Days.Qualified.To.Total.Days          -0.250513511
## Ratio.Days.Validated.To.Total.Days
## Opportunity.Number                         -0.05753584
## Elapsed.Days.In.Sales.Stage                 0.01582840
## Sales.Stage.Change.Count                   -0.01320735
## Total.Days.Identified.Through.Closing        0.12477671
## Total.Days.Identified.Through.Qualified      0.14574872
## Opportunity.Amount.USD                     -0.05182146
## Ratio.Days.Identified.To.Total.Days          -0.47666745
## Ratio.Days.Validated.To.Total.Days          1.00000000
## Ratio.Days.Qualified.To.Total.Days          -0.43473516
## Ratio.Days.Qualified.To.Total.Days
## Opportunity.Number                        -0.04840408
## Elapsed.Days.In.Sales.Stage                 0.02040566
## Sales.Stage.Change.Count                   0.28067095
## Total.Days.Identified.Through.Closing        0.02652307
## Total.Days.Identified.Through.Qualified      0.02808720
## Opportunity.Amount.USD                      0.04672572
## Ratio.Days.Identified.To.Total.Days          -0.25051351
## Ratio.Days.Validated.To.Total.Days          -0.43473516
## Ratio.Days.Qualified.To.Total.Days          1.00000000

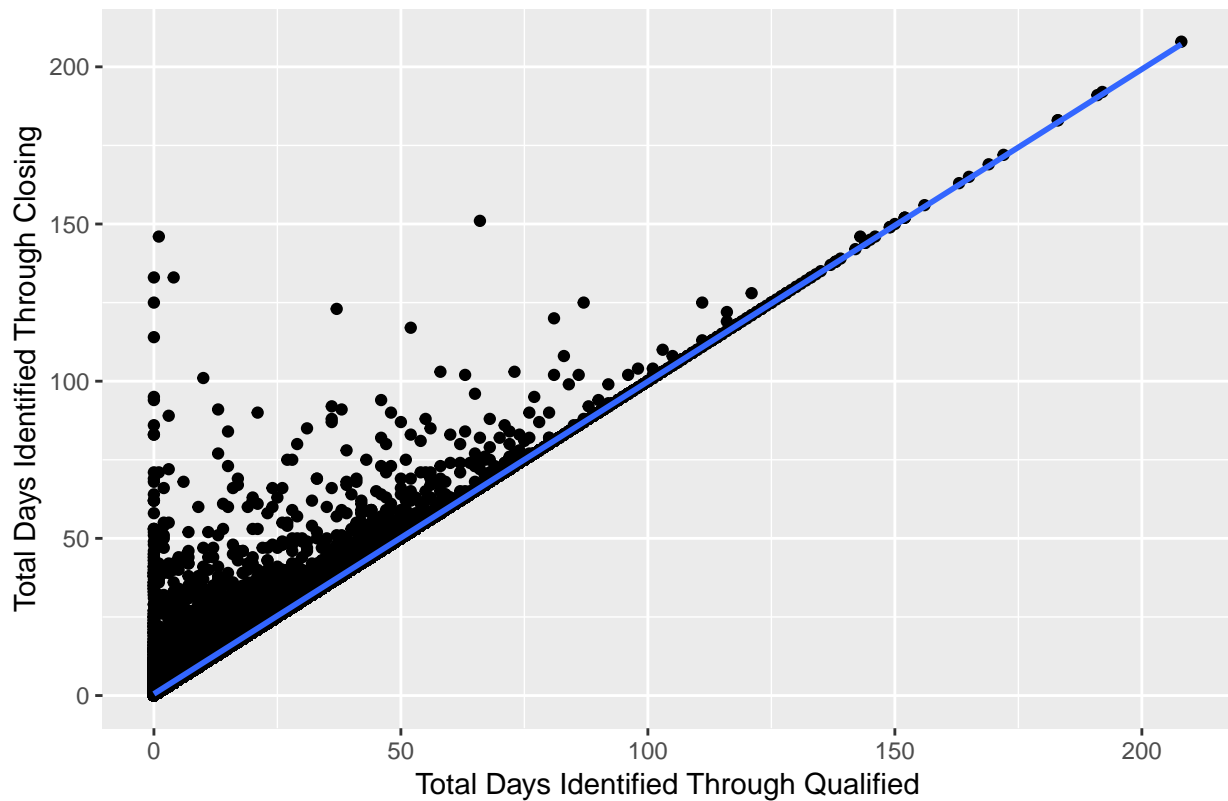
```

Let's have a look at the variables that are significantly correlated (say correlation coefficient either greater than 0.8 or less than -0.8).

We note that `Total.Days.Identified.Through.Qualified` and `Total.Days.Identified.Through.Closing` are strongly correlated (0.98), which is not surprising as these two variables are related in such that an opportunity stay in the pipeline from identification, through qualification and validation, to closing.

None of the other numeric features are strongly correlated.

Total Days Through Qualified vs Total Days Through Closing



```
##                               Supplies.Subgroup Supplies.Group
## Supplies.Subgroup              0.000000e+00  4.997501e-04
## Supplies.Group                  4.997501e-04  0.000000e+00
## Region                          0.000000e+00  1.838690e-65
## Route.To.Market                 4.997501e-04  4.997501e-04
## Client.Size.By.Revenue           2.366019e-119 1.658829e-04
## Client.Size.By.Employee.Count    2.549308e-94 5.520897e-01
## Revenue.From.Client.Past.Two.Years 4.997501e-04 4.997501e-04
## Competitor.Type                  4.915601e-283 1.561020e-05
## Deal.Size.Category                0.000000e+00 1.329050e-110
##                               Region Route.To.Market
## Supplies.Subgroup              0.000000e+00  4.997501e-04
## Supplies.Group                  1.838690e-65  4.997501e-04
## Region                          0.000000e+00  0.000000e+00
## Route.To.Market                 0.000000e+00  0.000000e+00
## Client.Size.By.Revenue           0.000000e+00  0.000000e+00
## Client.Size.By.Employee.Count    0.000000e+00 1.490446e-293
## Revenue.From.Client.Past.Two.Years 2.677537e-84 6.465001e-262
## Competitor.Type                  0.000000e+00  0.000000e+00
## Deal.Size.Category                1.490902e-171 0.000000e+00
##                               Client.Size.By.Revenue
## Supplies.Subgroup              2.366019e-119
## Supplies.Group                  1.658829e-04
## Region                          0.000000e+00
## Route.To.Market                 0.000000e+00
## Client.Size.By.Revenue           0.000000e+00
## Client.Size.By.Employee.Count    0.000000e+00
```

```

## Revenue.From.Client.Past.Two.Years      1.473843e-108
## Competitor.Type                          0.000000e+00
## Deal.Size.Category                       1.998643e-287
##                                           Client.Size.By.Employee.Count
## Supplies.Subgroup                       2.549308e-94
## Supplies.Group                           5.520897e-01
## Region                                  0.000000e+00
## Route.To.Market                         1.490446e-293
## Client.Size.By.Revenue                  0.000000e+00
## Client.Size.By.Employee.Count           0.000000e+00
## Revenue.From.Client.Past.Two.Years      5.760970e-76
## Competitor.Type                         0.000000e+00
## Deal.Size.Category                       3.398658e-271
##                                           Revenue.From.Client.Past.Two.Years
## Supplies.Subgroup                       4.997501e-04
## Supplies.Group                           4.997501e-04
## Region                                  2.677537e-84
## Route.To.Market                         6.465001e-262
## Client.Size.By.Revenue                  1.473843e-108
## Client.Size.By.Employee.Count           5.760970e-76
## Revenue.From.Client.Past.Two.Years      0.000000e+00
## Competitor.Type                         3.062589e-72
## Deal.Size.Category                       0.000000e+00
##                                           Competitor.Type Deal.Size.Category
## Supplies.Subgroup                       4.915601e-283 0.000000e+00
## Supplies.Group                           1.561020e-05 1.329050e-110
## Region                                  0.000000e+00 1.490902e-171
## Route.To.Market                         0.000000e+00 0.000000e+00
## Client.Size.By.Revenue                  0.000000e+00 1.998643e-287
## Client.Size.By.Employee.Count           0.000000e+00 3.398658e-271
## Revenue.From.Client.Past.Two.Years      3.062589e-72 0.000000e+00
## Competitor.Type                         0.000000e+00 0.000000e+00
## Deal.Size.Category                       0.000000e+00 0.000000e+00

```

Null hypothesis assumes that there is no association between two variables.

Here, we have all p-values < 0.05 , so we reject the null hypothesis and conclude that all the variables are dependent to each other.

In-depth data exploration for first insights

```

##
## Loss Won
## 60281 17548
##
## Loss Won
## 0.77 0.23

```

The success rate is quite low, **only 23% of our opportunities are converted into revenues.**

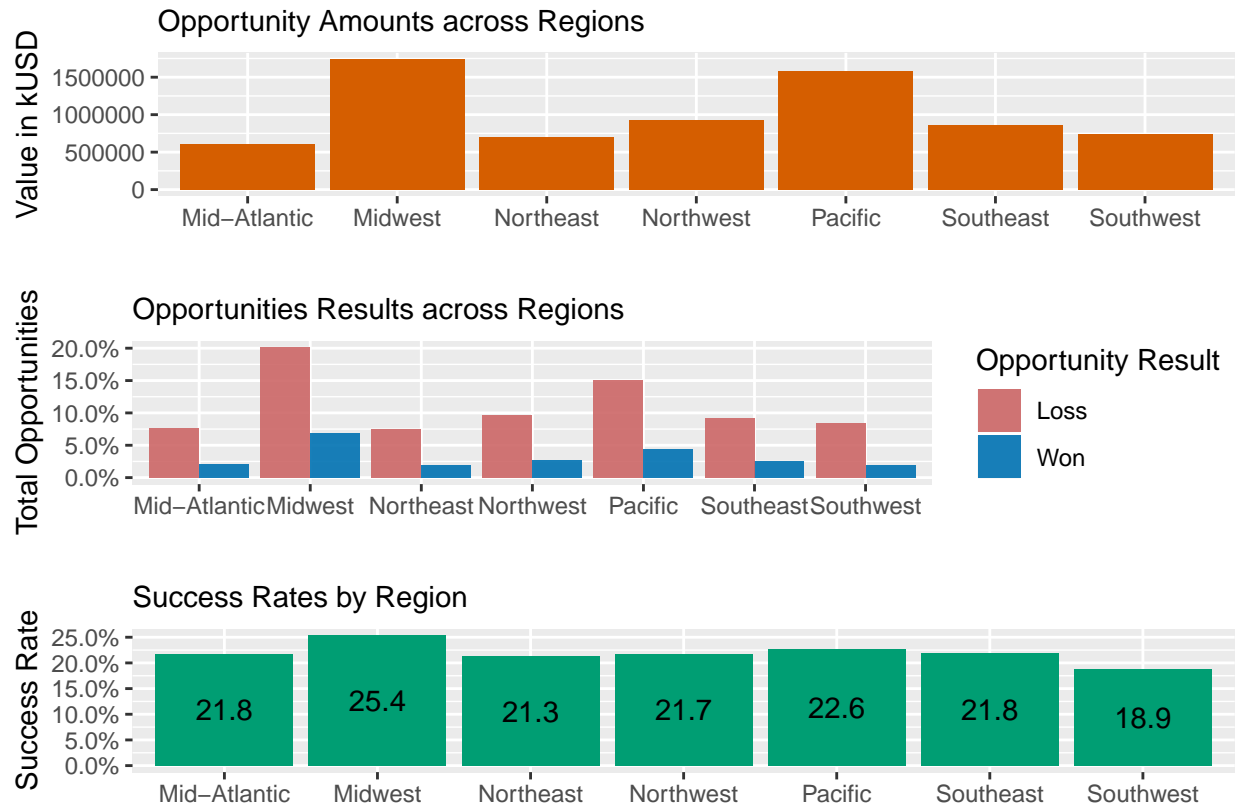
Now, as sales people, we rather focus on revenues and want to first check how we perform:

- across areas,
- by deal sizes,
- across sales channels.

Deal conversions accross areas

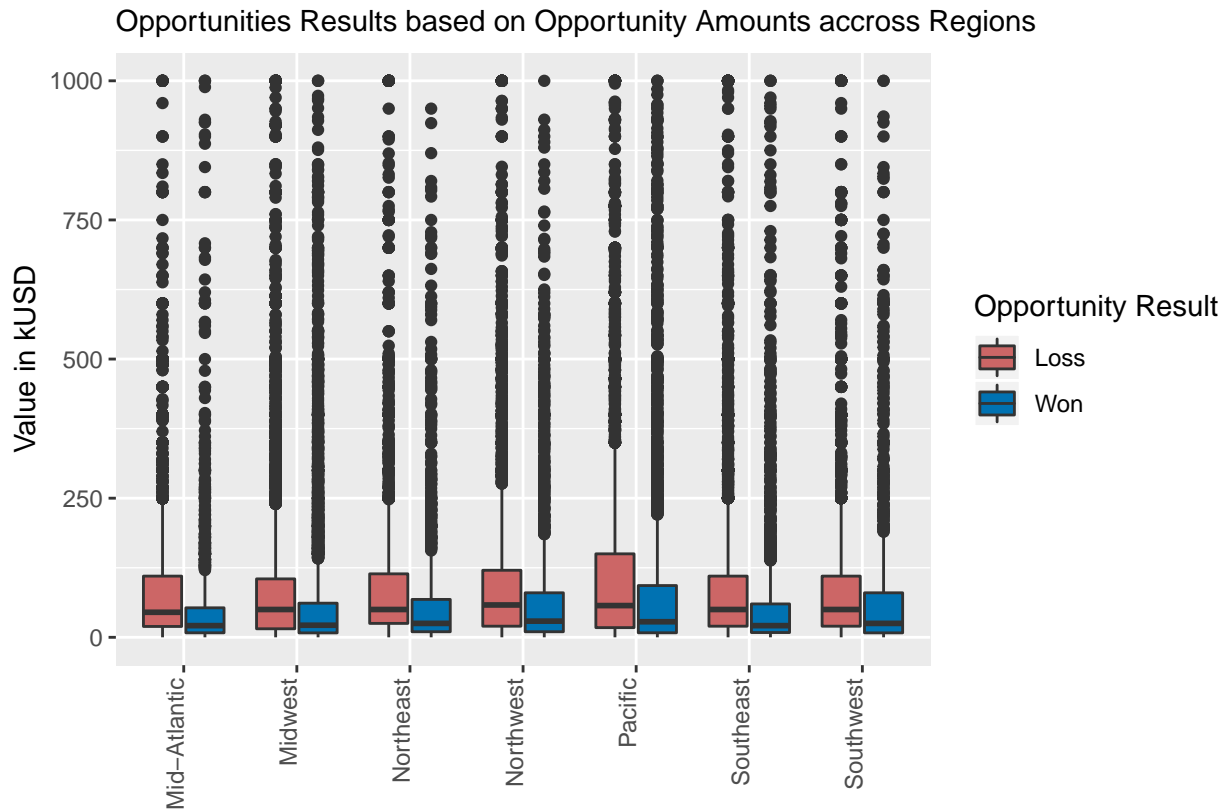
How do the Opportunity.Amount and Opportunity.Result compare by Region?

The maximum opportunity amount is 1000 thousand USD, the average is 91665 thousand USD, and the median is 60000 thousand USD.



Midwest and Pacific are our biggest areas in terms of opportunity amounts.

Our deal conversion rates across all regions are similarly low, so there is surely room for improving our sales efficiency!

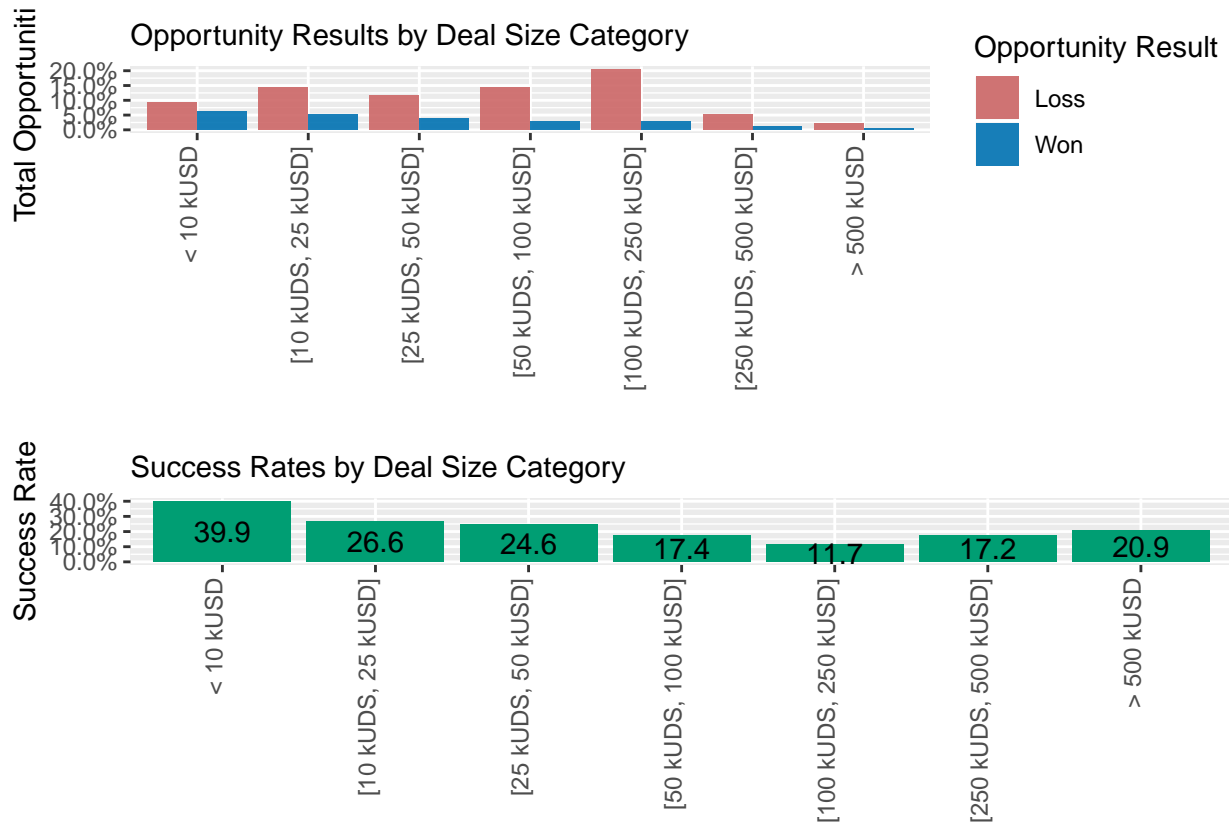


The majority of the opportunities are skewed on the low amounts. Interestingly, we note that the medians of won opportunities are with deals lower than 25 kUSD!

Let's see further how the `Opportunity.Result` compare by `Deal.Size.Category`.

Win / Loss opportunities by deal size categories

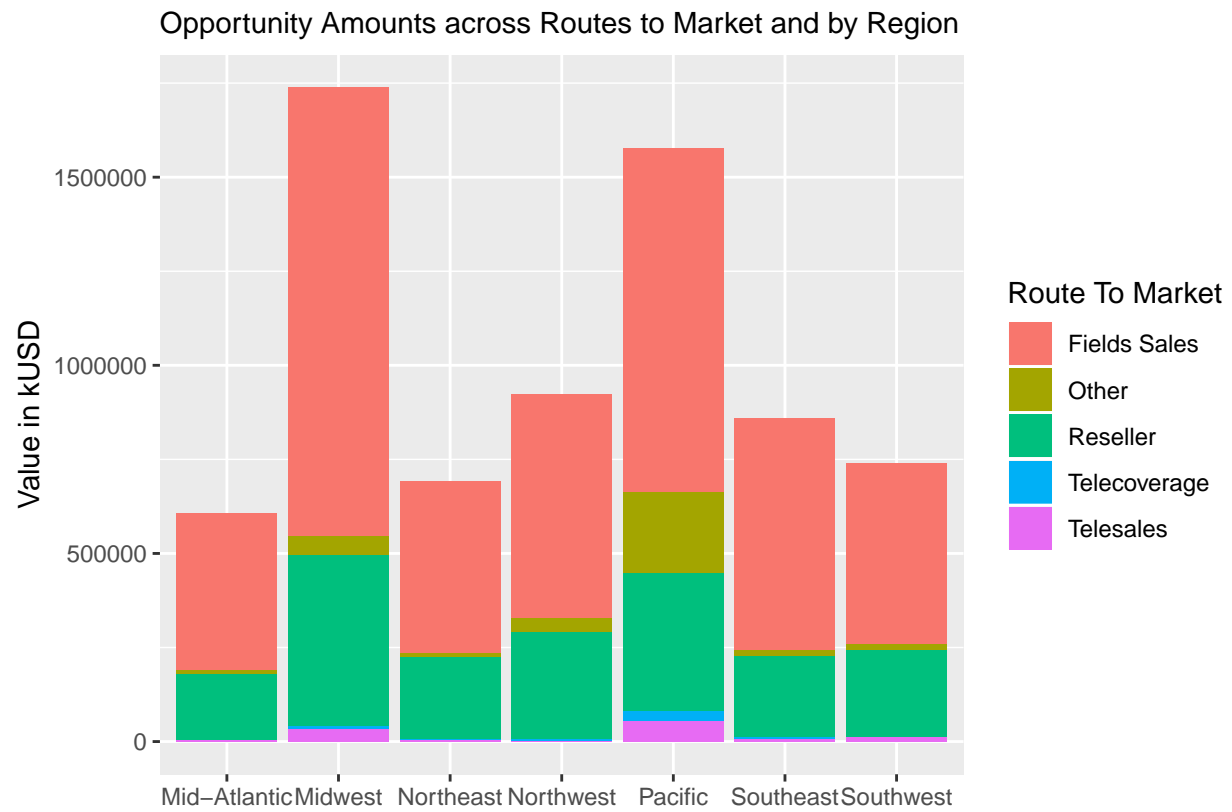
How do the `Opportunity.Result` compare by `Deal.Size.Category`?



We won more opportunities for deal size < 10 kUSD and our success rate decreases as the deal size increases until a level of 100 to 250 kUSD.

Opportunity amount accross regions and sales channels

How do the Opportunity.Amount compare by Region and Route.To.Market?

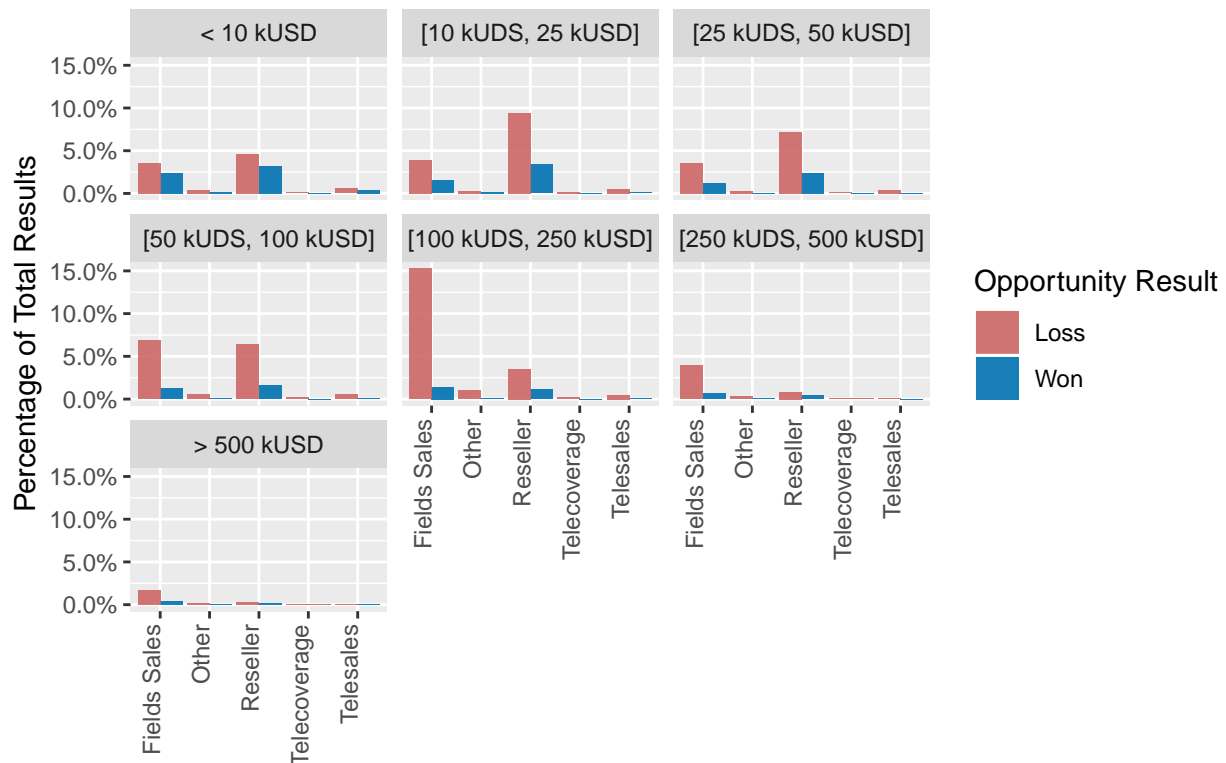


Field Sales and Reseller are our two main sales channels accross regions, but how do these sales channels perform?

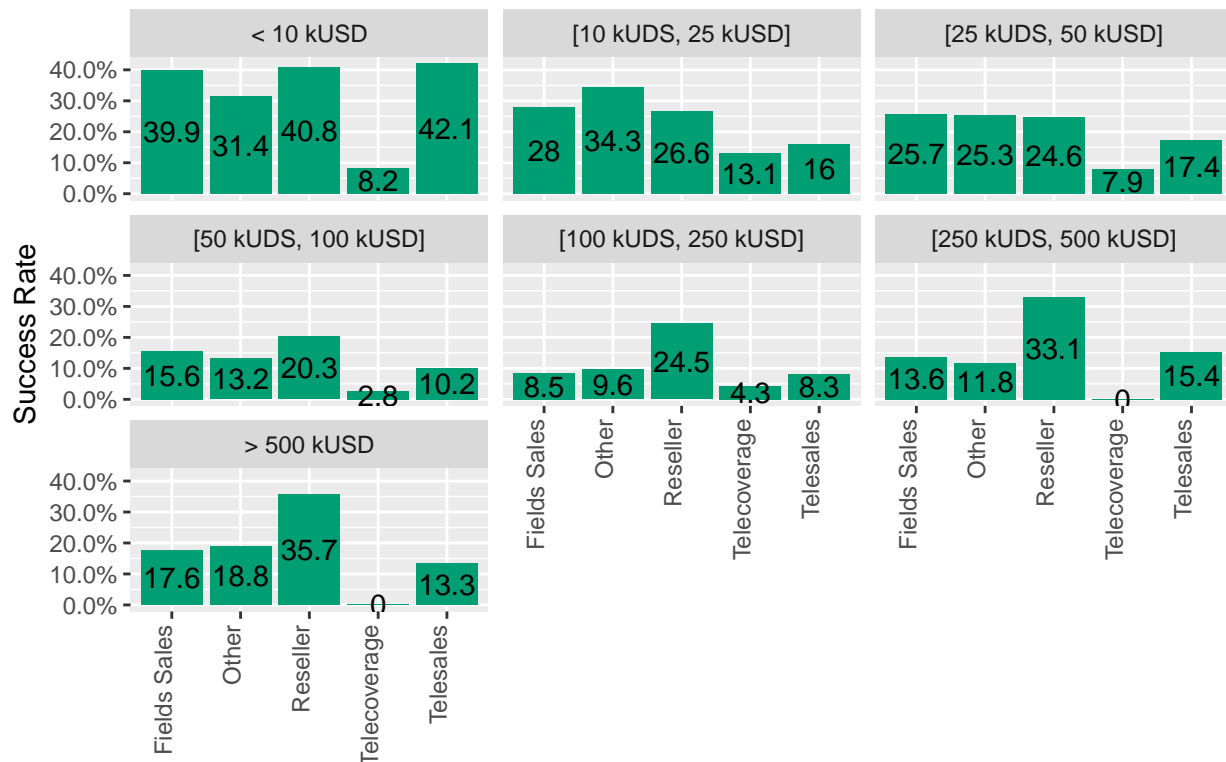
Win / Loss opportunities accross sales channels and by deal size category

How do the Opportunity.Result compare by Route.To.Market and Deal.Size.Category?

Opportunity Results by Route to Market and Deal Size Category



Success Rates by Route To Market and by Deal Size Category



There is no clear trend of successful routes to market across deal sizes.

- Field Sales performance trend follows the general trend across the deal size categories as seen before.

So, we note that slight rebound for opportunities above 250 kUSD. Field Sales could be refocused on bigger deals to increase efficiency.

- Reseller channel performs well and best for all deals > 50 kUSD. We should further develop and support our resellers network.
- Other routes to market efficiency could be somewhat compared to Field Sales.
- Telecoverage performance is poor, it might not be relevant for our business.
- Telesales is the best channel for deals < 10 kUSD.

Based on this finding, we may consider shifting our sales resources as:

- Telesales for opportunities < 10 kUSD,
- Field Sales for opportunities > 250 kUSD,
- Reseller for all opportunities,
- Other for all opportunities,
- Telecoverage should be discontinued.

Understanding what drives our sales

With a simple managerial approach, looking at a few variables (`Opportunity.Amount`, `Region`, `Deal.Size.Category` and `Route.To.Market`), we have not been able to uncover the patterns that allow us to determine the successful sales profiles. The best performance we could achieve was a modest 40% of deals conversion with Telesales for opportunities < 10 kUSD.

We want to understand what drives our sales, which deals our sales team can expect to win or loose. In other terms, we want to understand the **why** behind what's happening.

With such a large dataset including 19 variables (so as many as 18 possible sales drivers), we can't manually explore each and every variable, not even talking about possible combinations. That is where we bring in **Machine Learning approaches to help us to identify the most significant variables and predict the opportunity results.**

Methods to perform a dimension reduction of our dataset so that we can identify the most significant variables

Two very common methods for identifying significant variables are **Decision Tree** and **Random Forests**.

- The **Decision Tree** best feature for analytics is that it is very **easy to interpret** and **results are actionable!**
- **Random Forests** improve the robustness of our predictions as they aggregate many Decision Trees.

4.2 Create subsets for the project

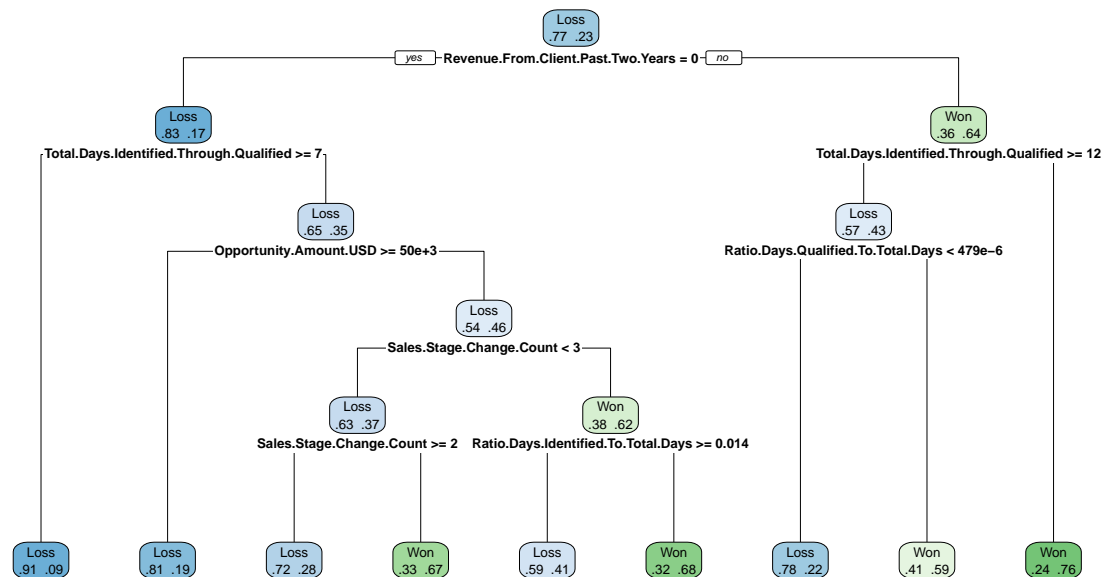
4.3 Predictive Models

Decision Tree

```
##
## Classification tree:
## rpart(formula = Opportunity.Result ~ . - Opportunity.Number,
##       data = train_set, method = "class")
##
```



```
## Variables actually used in tree construction:
## [1] Opportunity.Amount.USD
## [2] Ratio.Days.Identified.To.Total.Days
## [3] Ratio.Days.Qualified.To.Total.Days
## [4] Revenue.From.Client.Past.Two.Years
## [5] Sales.Stage.Change.Count
## [6] Total.Days.Identified.Through.Qualified
##
## Root node error: 12634/56035 = 0.22547
##
## n= 56035
##
##      CP nsplit rel error  xerror    xstd
## 1 0.140019      0  1.00000 1.00000 0.0078298
## 2 0.025566      1  0.85998 0.85998 0.0074075
## 3 0.021002      2  0.83442 0.84439 0.0073560
## 4 0.019550      6  0.73468 0.74861 0.0070180
## 5 0.011477      7  0.71513 0.71545 0.0068916
## 6 0.010000      8  0.70366 0.70659 0.0068569
```



```
## Call:
## rpart(formula = Opportunity.Result ~ . - Opportunity.Number,
##       data = train_set, method = "class")
##   n= 56035
##
##      CP nsplit rel error  xerror    xstd
## 1 0.14001900      0 1.0000000 1.0000000 0.007829783
## 2 0.02556593      1  0.8599810 0.8599810 0.007407461
## 3 0.02100153      2  0.8344151 0.8443882 0.007355987
## 4 0.01955042      6  0.7346842 0.7486148 0.007018024
## 5 0.01147697      7  0.7151338 0.7154504 0.006891600
## 6 0.01000000      8  0.7036568 0.7065854 0.006856927
##
## Variable importance
##   Revenue.From.Client.Past.Two.Years
##                                     29
```

```

## Total.Days.Identified.Through.Qualified
##                               21
##   Total.Days.Identified.Through.Closing
##                               21
##           Sales.Stage.Change.Count
##                               10
##           Opportunity.Amount.USD
##                               7
##   Ratio.Days.Qualified.To.Total.Days
##                               3
##   Ratio.Days.Identified.To.Total.Days
##                               3
##   Ratio.Days.Validated.To.Total.Days
##                               3
##           Supplies.Subgroup
##                               1
##           Route.To.Market
##                               1
##           Competitor.Type
##                               1
##
## Node number 1: 56035 observations,   complexity param=0.140019
##   predicted class=Loss   expected loss=0.2254662   P(node) =1
##   class counts: 43401 12634
##   probabilities: 0.775 0.225
##   left son=2 (49730 obs) right son=3 (6305 obs)
##   Primary splits:
##       Revenue.From.Client.Past.Two.Years      splits as   LRRRR, improve=2444.9720, (0 missing)
##       Total.Days.Identified.Through.Qualified < 6.5      to the right, improve=2157.3690, (0 missing)
##       Total.Days.Identified.Through.Closing   < 8.5      to the right, improve=1903.3660, (0 missing)
##       Ratio.Days.Qualified.To.Total.Days      < 0.0167765 to the left,  improve=1161.7860, (0 missing)
##       Sales.Stage.Change.Count                < 1.5      to the right, improve= 884.6567, (0 missing)
##   Surrogate splits:
##       Total.Days.Identified.Through.Closing   < 187      to the left,  agree=0.887, adj=0, (0 split)
##       Total.Days.Identified.Through.Qualified < 187      to the left,  agree=0.887, adj=0, (0 split)
##
## Node number 2: 49730 observations,   complexity param=0.02100153
##   predicted class=Loss   expected loss=0.1728735   P(node) =0.887481
##   class counts: 41133 8597
##   probabilities: 0.827 0.173
##   left son=4 (33535 obs) right son=5 (16195 obs)
##   Primary splits:
##       Total.Days.Identified.Through.Qualified < 6.5      to the right, improve=1462.8840, (0 missing)
##       Total.Days.Identified.Through.Closing   < 6.5      to the right, improve=1297.2330, (0 missing)
##       Sales.Stage.Change.Count                < 1.5      to the right, improve= 706.7389, (0 missing)
##       Ratio.Days.Qualified.To.Total.Days      < 0.0167765 to the left,  improve= 596.8093, (0 missing)
##       Opportunity.Amount.USD                 < 19997     to the right, improve= 571.3677, (0 missing)
##   Surrogate splits:
##       Total.Days.Identified.Through.Closing   < 6.5      to the right, agree=0.989, adj=0.966, (0 split)
##       Sales.Stage.Change.Count                < 1.5      to the right, agree=0.715, adj=0.125, (0 split)
##       Ratio.Days.Validated.To.Total.Days      < 0.0007405 to the right, agree=0.697, adj=0.069, (0 split)
##       Opportunity.Amount.USD                 < 2962.5    to the right, agree=0.680, adj=0.019, (0 split)
##       Ratio.Days.Qualified.To.Total.Days      < 0.9996015 to the left,  agree=0.679, adj=0.015, (0 split)
##

```

```

## Node number 3: 6305 observations,      complexity param=0.02556593
##   predicted class=Won   expected loss=0.3597145   P(node) =0.112519
##   class counts:  2268  4037
##   probabilities:  0.360  0.640
##   left son=6 (2321 obs) right son=7 (3984 obs)
##   Primary splits:
##     Total.Days.Identified.Through.Qualified < 11.5      to the right, improve=323.5650, (0 missing)
##     Total.Days.Identified.Through.Closing   < 11.5      to the right, improve=277.2564, (0 missing)
##     Opportunity.Amount.USD                  < 98842      to the right, improve=252.2435, (0 missing)
##     Deal.Size.Category                      splits as  RRRRLLL, improve=251.5352, (0 missing)
##     Ratio.Days.Identified.To.Total.Days     < 0.0022035 to the right, improve=209.8633, (0 missing)
##   Surrogate splits:
##     Total.Days.Identified.Through.Closing < 11.5      to the right, agree=0.980, adj=0.944, (0 spl
##     Ratio.Days.Identified.To.Total.Days  < 0.0009805 to the right, agree=0.688, adj=0.152, (0 spl
##     Sales.Stage.Change.Count              < 4.5         to the right, agree=0.677, adj=0.123, (0 spl
##     Opportunity.Amount.USD                < 94972.5     to the right, agree=0.647, adj=0.041, (0 spl
##     Competitor.Type                      splits as  LRR, agree=0.647, adj=0.041, (0 split)
##
## Node number 4: 33535 observations
##   predicted class=Loss expected loss=0.08859401   P(node) =0.5984652
##   class counts:  30564  2971
##   probabilities:  0.911  0.089
##
## Node number 5: 16195 observations,      complexity param=0.02100153
##   predicted class=Loss expected loss=0.3473912   P(node) =0.2890158
##   class counts:  10569  5626
##   probabilities:  0.653  0.347
##   left son=10 (6754 obs) right son=11 (9441 obs)
##   Primary splits:
##     Opportunity.Amount.USD                  < 49967.5   to the right, improve=568.8962, (0 missing)
##     Deal.Size.Category                      splits as  RRRRLLL, improve=568.8962, (0 missing)
##     Route.To.Market                        splits as  LLRLL, improve=502.3970, (0 missing)
##     Sales.Stage.Change.Count                < 2.5         to the left,  improve=462.8220, (0 missing)
##     Ratio.Days.Qualified.To.Total.Days     < 0.004385   to the left,  improve=348.3002, (0 missing)
##   Surrogate splits:
##     Supplies.Subgroup                      splits as  RRRRLRLRLLL, agree=0.625, adj=0.101, (0 split)
##     Route.To.Market                        splits as  LRRLR, agree=0.621, adj=0.090, (0 split)
##     Ratio.Days.Identified.To.Total.Days < 0.145286   to the right, agree=0.612, adj=0.070, (0 split)
##     Competitor.Type                      splits as  LRR, agree=0.604, adj=0.051, (0 split)
##     Client.Size.By.Employee.Count          splits as  RRRRL, agree=0.588, adj=0.012, (0 split)
##
## Node number 6: 2321 observations,      complexity param=0.01955042
##   predicted class=Loss expected loss=0.4304179   P(node) =0.04142054
##   class counts:  1322   999
##   probabilities:  0.570  0.430
##   left son=12 (1016 obs) right son=13 (1305 obs)
##   Primary splits:
##     Ratio.Days.Qualified.To.Total.Days < 0.000479   to the left,  improve=160.79180, (0 missing)
##     Sales.Stage.Change.Count            < 3.5         to the left,  improve=103.11550, (0 missing)
##     Ratio.Days.Identified.To.Total.Days < 0.9794605   to the right, improve= 60.33883, (0 missing)
##     Revenue.From.Client.Past.Two.Years splits as  -RRL, improve= 51.42239, (0 missing)
##     Opportunity.Amount.USD              < 71182       to the right, improve= 50.62278, (0 missing)
##   Surrogate splits:
##     Ratio.Days.Validated.To.Total.Days < 0.871351   to the right, agree=0.789, adj=0.519, (0 spl

```

```

##      Sales.Stage.Change.Count          < 2.5      to the left,  agree=0.762, adj=0.457, (0 spl
##      Ratio.Days.Identified.To.Total.Days < 0.881062 to the right, agree=0.660, adj=0.223, (0 spl
##      Opportunity.Amount.USD             < 11921    to the left,  agree=0.569, adj=0.016, (0 spl
##      Total.Days.Identified.Through.Closing < 14.5     to the left,  agree=0.566, adj=0.009, (0 spl
##
## Node number 7: 3984 observations
##   predicted class=Won   expected loss=0.2374498  P(node) =0.07109842
##   class counts:   946   3038
##   probabilities: 0.237 0.763
##
## Node number 10: 6754 observations
##   predicted class=Loss expected loss=0.1907018  P(node) =0.1205318
##   class counts:   5466  1288
##   probabilities: 0.809 0.191
##
## Node number 11: 9441 observations,   complexity param=0.02100153
##   predicted class=Loss expected loss=0.4594852  P(node) =0.168484
##   class counts:   5103  4338
##   probabilities: 0.541 0.459
##   left son=22 (6029 obs) right son=23 (3412 obs)
##   Primary splits:
##     Sales.Stage.Change.Count          < 2.5      to the left,  improve=263.9407, (0 missing)
##     Ratio.Days.Qualified.To.Total.Days < 0.016287 to the left,  improve=182.9088, (0 missing)
##     Ratio.Days.Identified.To.Total.Days < 0.9622505 to the right, improve=132.9027, (0 missing)
##     Route.To.Market                   splits as  LLRLL, improve=131.7445, (0 missing)
##     Elapsed.Days.In.Sales.Stage       < 91.5     to the left,  improve=112.6910, (0 missing)
##   Surrogate splits:
##     Ratio.Days.Qualified.To.Total.Days < 0.0007585 to the left,  agree=0.767, adj=0.356, (0 spl
##     Ratio.Days.Validated.To.Total.Days < 0.0003995 to the left,  agree=0.685, adj=0.129, (0 spl
##     Total.Days.Identified.Through.Closing < 6.5      to the left,  agree=0.672, adj=0.091, (0 spl
##     Ratio.Days.Identified.To.Total.Days < 0.0015175 to the left,  agree=0.658, adj=0.055, (0 spl
##     Competitor.Type                   splits as  RLL, agree=0.648, adj=0.025, (0 split)
##
## Node number 12: 1016 observations
##   predicted class=Loss expected loss=0.2194882  P(node) =0.01813152
##   class counts:   793   223
##   probabilities: 0.781 0.219
##
## Node number 13: 1305 observations
##   predicted class=Won   expected loss=0.405364  P(node) =0.02328902
##   class counts:   529   776
##   probabilities: 0.405 0.595
##
## Node number 22: 6029 observations,   complexity param=0.02100153
##   predicted class=Loss expected loss=0.3705424  P(node) =0.1075935
##   class counts:   3795  2234
##   probabilities: 0.629 0.371
##   left son=44 (4643 obs) right son=45 (1386 obs)
##   Primary splits:
##     Sales.Stage.Change.Count          < 1.5      to the right, improve=317.17690, (0 missing)
##     Route.To.Market                   splits as  LLRLL, improve=127.59050, (0 missing)
##     Elapsed.Days.In.Sales.Stage       < 91.5     to the left,  improve= 92.07853, (0 missing)
##     Opportunity.Amount.USD             < 2.5      to the left,  improve= 89.70116, (0 missing)
##     Total.Days.Identified.Through.Qualified < 0.5     to the right, improve= 86.47074, (0 missing)

```

```

## Surrogate splits:
##   Elapsed.Days.In.Sales.Stage < 93.5      to the left,  agree=0.771, adj=0.006, (0 split)
##   Opportunity.Amount.USD      < 49502      to the left,  agree=0.771, adj=0.003, (0 split)
##
## Node number 23: 3412 observations,      complexity param=0.01147697
##   predicted class=Won   expected loss=0.3833529   P(node) =0.06089051
##   class counts:  1308   2104
##   probabilities:  0.383 0.617
##   left son=46 (791 obs) right son=47 (2621 obs)
##   Primary splits:
##     Ratio.Days.Identified.To.Total.Days      < 0.0137925 to the right, improve=89.35948, (0 missing)
##     Total.Days.Identified.Through.Qualified < 2.5          to the right, improve=48.82247, (0 missing)
##     Total.Days.Identified.Through.Closing  < 4.5          to the right, improve=33.91275, (0 missing)
##     Ratio.Days.Qualified.To.Total.Days      < 0.0413815 to the left,  improve=30.10667, (0 missing)
##     Opportunity.Amount.USD                  < 9999.5      to the right, improve=22.64671, (0 missing)
##   Surrogate splits:
##     Supplies.Subgroup      splits as RRRRRRRRRLR, agree=0.768, adj=0.001, (0 split)
##     Supplies.Group         splits as RRRRL, agree=0.768, adj=0.001, (0 split)
##     Sales.Stage.Change.Count < 8.5          to the right, agree=0.768, adj=0.001, (0 split)
##
## Node number 44: 4643 observations
##   predicted class=Loss expected loss=0.2819298   P(node) =0.08285893
##   class counts:  3334   1309
##   probabilities:  0.718 0.282
##
## Node number 45: 1386 observations
##   predicted class=Won   expected loss=0.3326118   P(node) =0.02473454
##   class counts:  461    925
##   probabilities:  0.333 0.667
##
## Node number 46: 791 observations
##   predicted class=Loss expected loss=0.4083439   P(node) =0.01411618
##   class counts:  468    323
##   probabilities:  0.592 0.408
##
## Node number 47: 2621 observations
##   predicted class=Won   expected loss=0.3204884   P(node) =0.04677434
##   class counts:  840   1781
##   probabilities:  0.320 0.680
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  Loss   Won
##   Loss 10132  1555
##   Won   719   1604
##
##           Accuracy : 0.8377
##           95% CI : (0.8315, 0.8438)
##   No Information Rate : 0.7745
##   P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.4872
##   McNemar's Test P-Value : < 2.2e-16

```

```
##
##          Sensitivity : 0.9337
##          Specificity : 0.5078
##          Pos Pred Value : 0.8669
##          Neg Pred Value : 0.6905
##          Prevalence : 0.7745
##          Detection Rate : 0.7232
##          Detection Prevalence : 0.8342
##          Balanced Accuracy : 0.7207
##
##          'Positive' Class : Loss
##
```

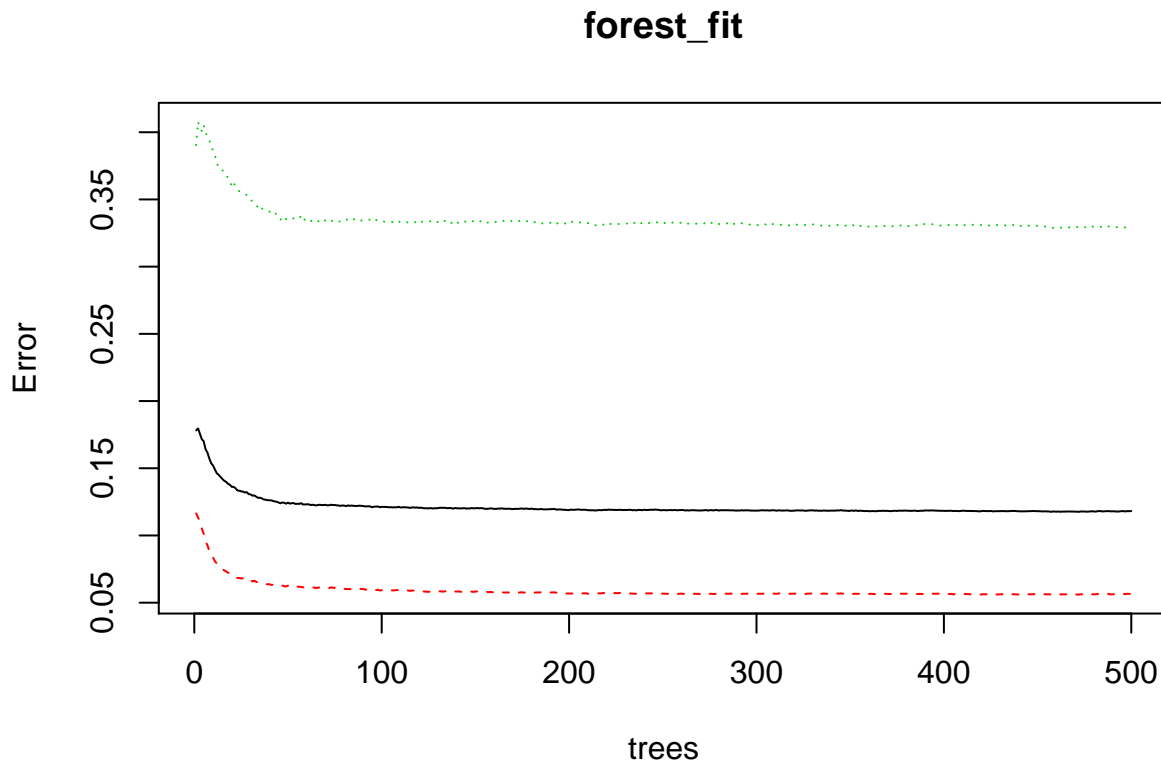
We have achieved a decent overall accuracy of 84% with a negative predictive value (the proportion of predicted won opportunities which are real won deals) of 69%.

More importantly, the Decision Tree helped us to understand not only the relationships and associations between features but also the decision rules to generate that tree.

So the feature of first importance is `Revenue.From.Client.Past.Two.Years`, i.e. the business that we've had with that Customer during the past two years, then the second feature is `Total.Days.Identified.Through.Qualified`, i.e. the number of days to qualify an opportunity from its identification. We may note that `Total.Days.Identified.Through.Closing` can be used as a second feature too. These two variables are strongly correlated as we have seen before and basically bear the same information. The third significant feature is `Sales.Stage.Change.Count`, i.e. the number of times an opportunity changes sales stages (back and forwards) in the sales pipeline.

We will focus on two or three features only as we want to keep our insights interpretable and above all actionable.

Random Forest



The green, black and red lines represent error rate for Loss, overall and Won, respectively. The overall error rate converges (no further decrease) to around 12%, so the default setting of 500 trees in the randomForest function is fine.

```
##                               MeanDecreaseGini
## Revenue.From.Client.Past.Two.Years      2026.8183
## Total.Days.Identified.Through.Qualified  2014.5458
## Elapsed.Days.In.Sales.Stage             1985.6999
## Opportunity.Amount.USD                  1917.6508
## Total.Days.Identified.Through.Closing    1671.7124
## Sales.Stage.Change.Count                 1402.7043
## Ratio.Days.Qualified.To.Total.Days       1324.6450
## Supplies.Subgroup                       1123.4961
## Region                                  1070.7206
## Ratio.Days.Validated.To.Total.Days       917.7806
## Deal.Size.Category                       733.1814
## Route.To.Market                         684.2728
## Ratio.Days.Identified.To.Total.Days      683.1925
## Client.Size.By.Employee.Count            372.3142
## Client.Size.By.Revenue                   366.5809
## Competitor.Type                         336.6062
## Supplies.Group                          157.0041

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  Loss   Won
##      Loss 10190  1021
##      Won   661  2138
##
##           Accuracy : 0.8799
##           95% CI : (0.8744, 0.8853)
##      No Information Rate : 0.7745
##      P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.6418
##  McNemar's Test P-Value : < 2.2e-16
##
##           Sensitivity : 0.9391
##           Specificity : 0.6768
##      Pos Pred Value : 0.9089
##      Neg Pred Value : 0.7638
##           Prevalence : 0.7745
##      Detection Rate : 0.7273
##      Detection Prevalence : 0.8002
##      Balanced Accuracy : 0.8079
##
##      'Positive' Class : Loss
##
```

With a Random Forest model, we have improved our overall accuracy to 88% with a negative predictive value of 76%.

Our top 3 predictors are:

- Revenue.From.Client.Past.Two.Years,

- Total.Days.Identified.Through.Qualified,
- Elapsed.Days.In.Sales.Stage, i.e. the number of days between the change in sales stages (the counter is reset for each new sales stage).

Note: The two first predictors are the same as given by the Decision Tree.

Validation

```
##                                     MeanDecreaseGini
## Total.Days.Identified.Through.Qualified      2576.4643
## Revenue.From.Client.Past.Two.Years          2544.4756
## Elapsed.Days.In.Sales.Stage                 2471.0120
## Opportunity.Amount.USD                     2447.5164
## Total.Days.Identified.Through.Closing        2014.2339
## Sales.Stage.Change.Count                   1702.8337
## Ratio.Days.Qualified.To.Total.Days          1645.0344
## Supplies.Subgroup                          1380.1869
## Region                                      1319.9473
## Ratio.Days.Validated.To.Total.Days          1144.1849
## Deal.Size.Category                          873.2417
## Ratio.Days.Identified.To.Total.Days          858.3467
## Route.To.Market                            856.7560
## Client.Size.By.Revenue                     459.6488
## Client.Size.By.Employee.Count              459.3147
## Competitor.Type                           430.4977
## Supplies.Group                             198.4108

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Loss  Won
##           Loss 5680  547
##           Won   349 1208
##
##           Accuracy : 0.8849
##           95% CI : (0.8776, 0.8919)
##           No Information Rate : 0.7745
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.6567
##           McNemar's Test P-Value : 4.663e-11
##
##           Sensitivity : 0.9421
##           Specificity : 0.6883
##           Pos Pred Value : 0.9122
##           Neg Pred Value : 0.7759
##           Prevalence : 0.7745
##           Detection Rate : 0.7297
##           Detection Prevalence : 0.8000
##           Balanced Accuracy : 0.8152
##
##           'Positive' Class : Loss
##
```

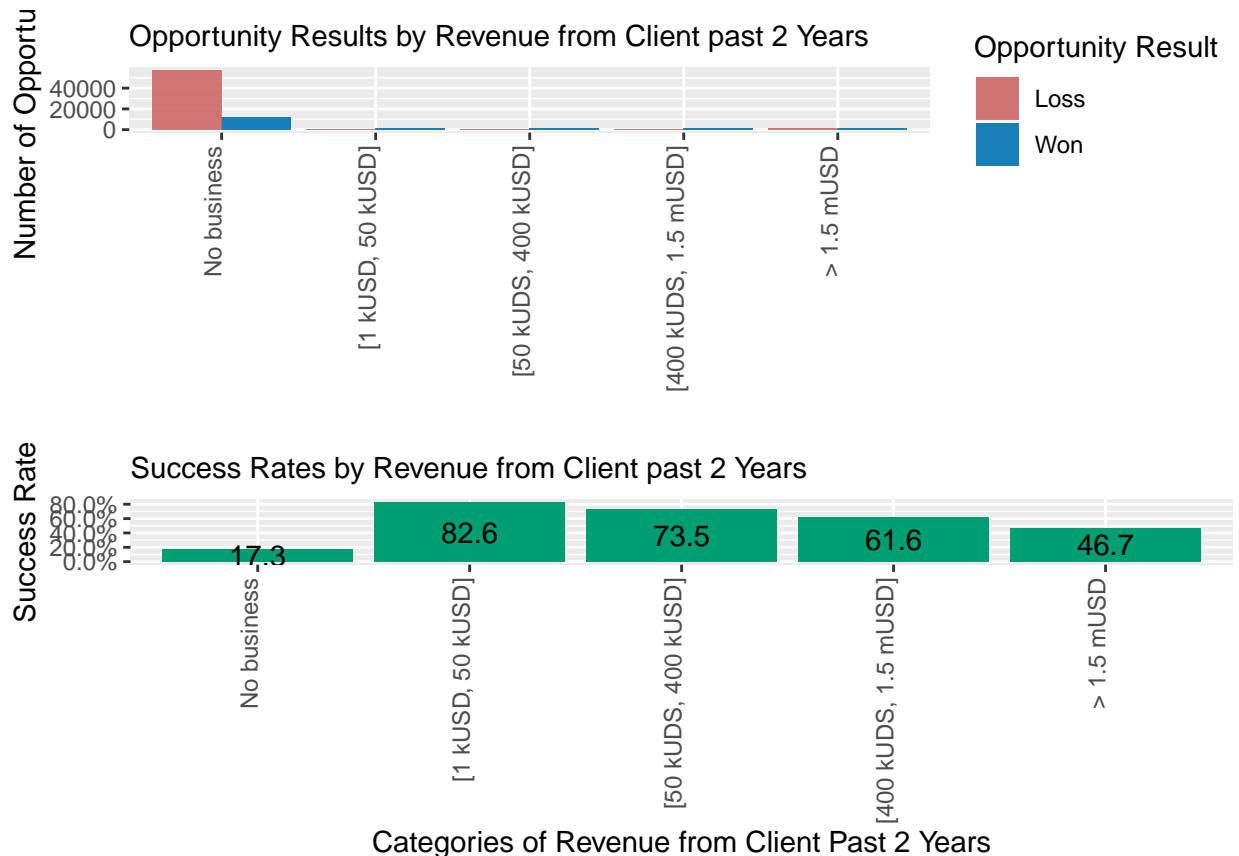

So we valid an overall accuracy of 88% with a negative predictive value of 77%.

Note that we didn't try to optimize accuracy by tuning our models, as our main goal was to reduce the dimension of our dataset and identify the most significant variables with their predictive strengths.

Let's see the insights we can gain from our predictive model.

Insights With One Predictor

We use the most significant predictor identified by our models, i.e. `Revenue.From.Client.Past.Two.Years`.



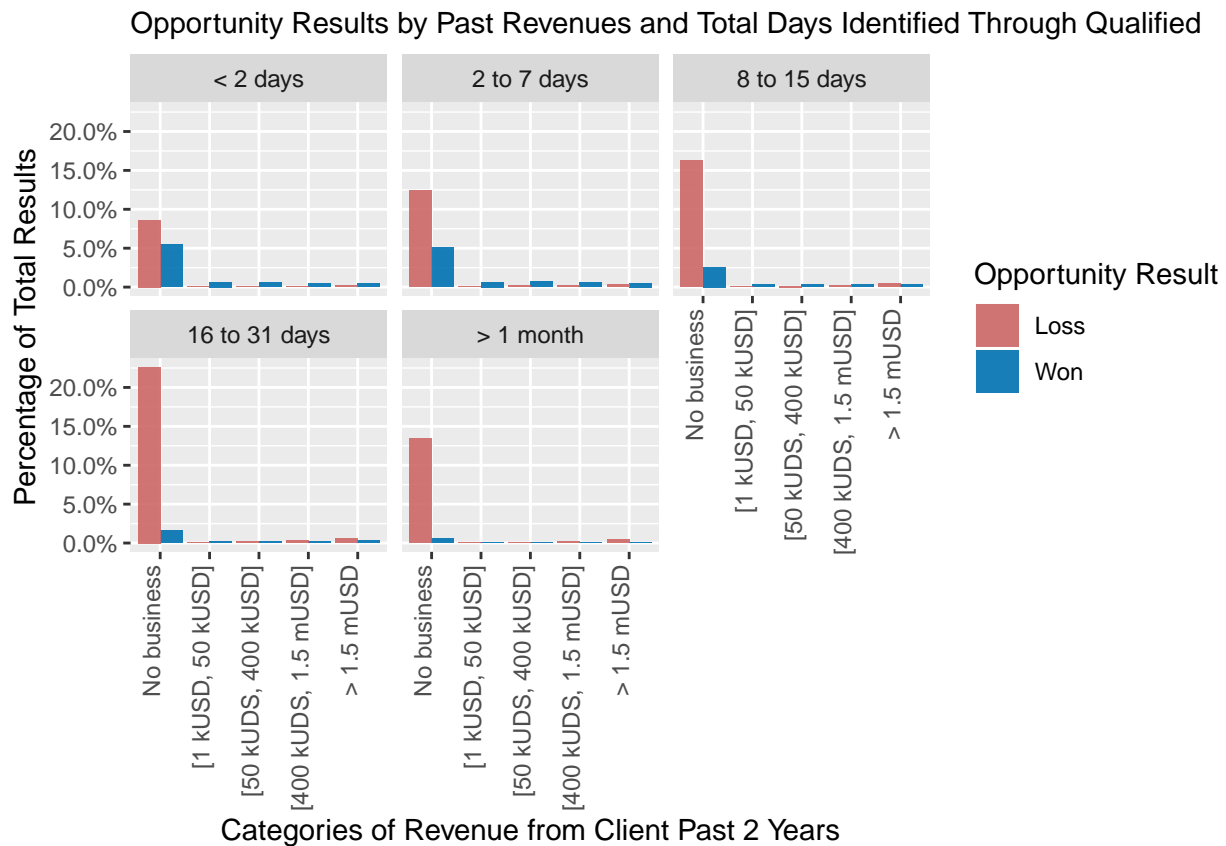
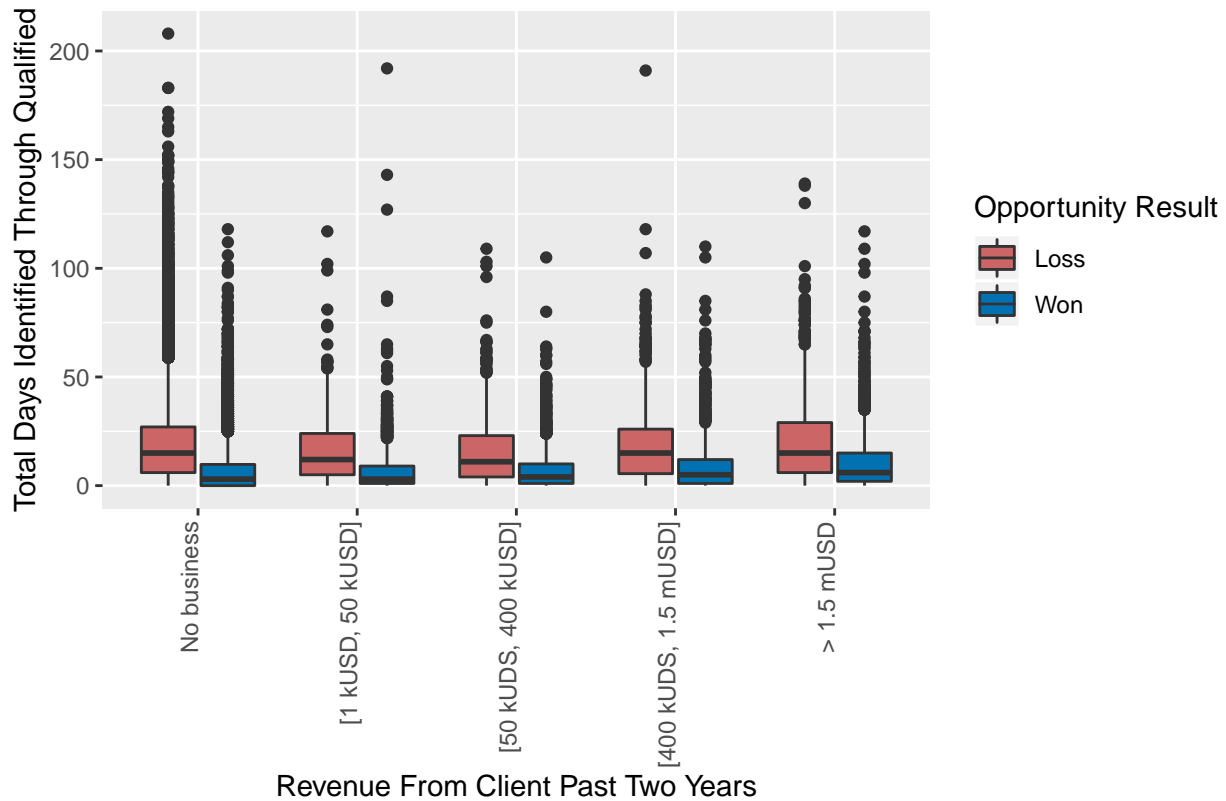
Looking at the Client purchase history, if they have bought from us less than 50,000 USD in the past 2 years, we have an **83% chance to successfully close the deal**.

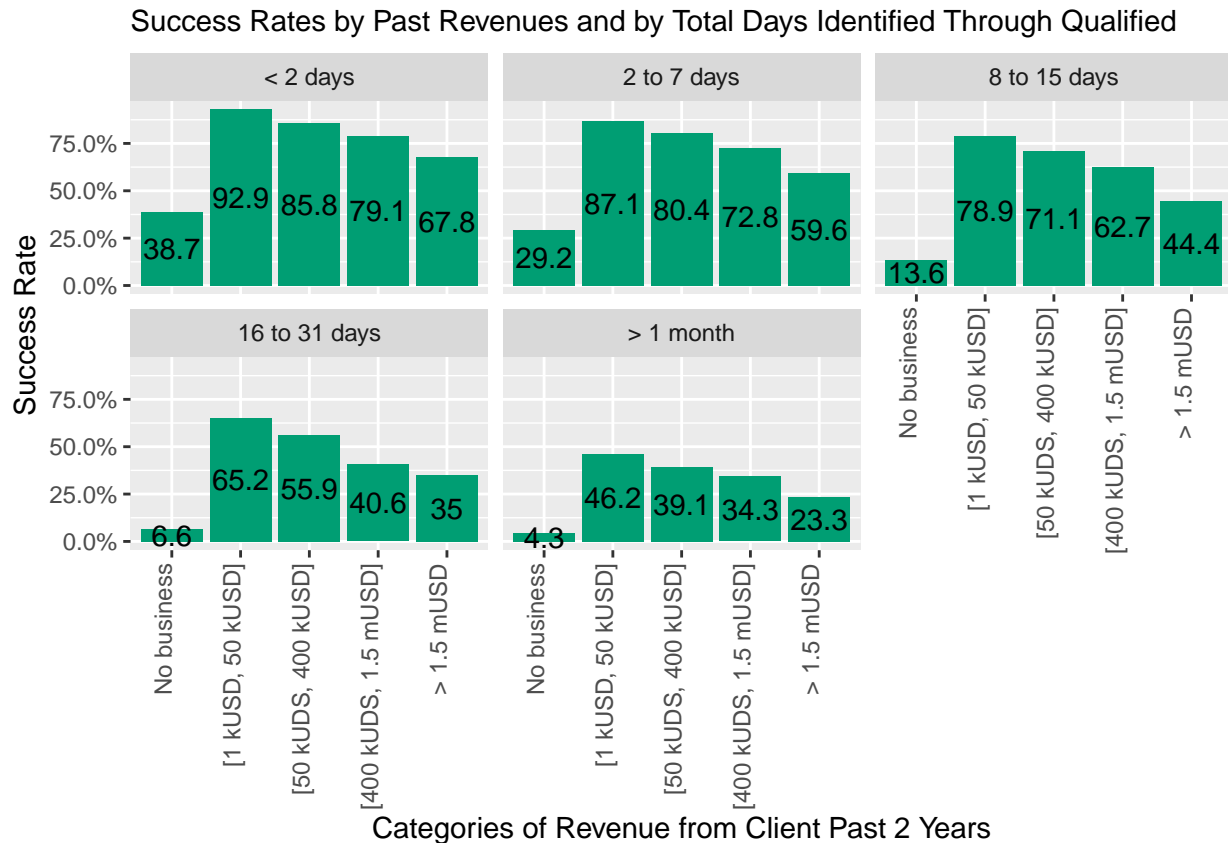
When it comes to very big opportunities (1.5 mUSD), we close the deal half the time. On the other end, gaining new customers is a real challenge with a success rate of only 17%.

Insights With Two Predictors

We use the two most important predictors identified by our models, i.e. `Revenue.From.Client.Past.Two.Years` and `Total.Days.Identified.Through.Qualified`.

Opportunities based on Total Days Identified Through Qualified and by Revenue From Client





If we can qualify, within 2 days, an opportunity with customers having purchased for less than 50,000 USD in the last 2 years, we reach **a probability of 93% to successfully close the deal**.

As a general rule, **the chances of winning a deal decreases as it stays longer in the pipeline**. This could help to formulate thresholds based on how many days a deal is in a pipeline and create alert mechanisms to expedite qualification.

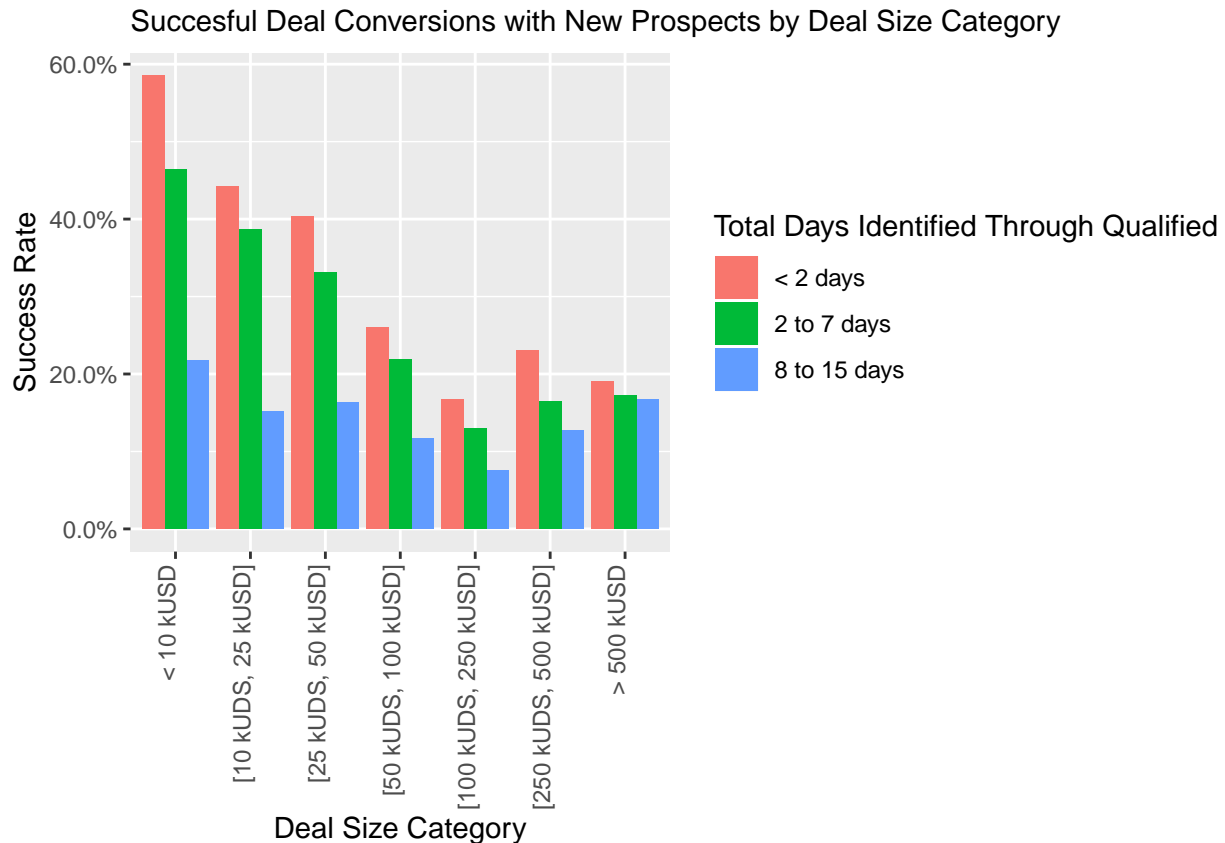
We also see the **same decrease trend with the increase of purchase history value, for a given qualification time frame**. For example, with an opportunity qualification of 2 to 7 days, we have an 87% chance of successful deal with customers valued at less than 50,000 USD and 60% with those at more than 1.5 mUSD.

We may also note that **an opportunity is more likely to result in a loss if the client didn't buy anything from us within the last 2 years** but if we are able to qualify a deal within a week with a new customer, we have more chance of success than our global (over the whole dataset) rate of 23%, as seen in the beginning of our analysis.

Insights With Three Predictors

We may want to uncover more complex relationships by adding more features, for example with three predictors as `Revenue.From.Client.Past.Two.Years`, `Total.Days.Identified.Through.Qualified` and `Opportunity.Amount.USD`.

Let's say we want to know how we perform with new prospects, with whose we managed to qualify the opportunities, regardless of the USD value, within two weeks.



As we know by now, the chances of successfully closing a deal are higher if we expedite the qualification of the opportunity. Nevertheless, this finding is balanced by the size of the deal. For higher value deals, the impact of qualification speed show a more nuanced picture.

More predictors?

At this stage, adding more predictors won't help much but may degrade the interpretability of our findings. Let's keep in mind that we want, above all, **insights that are relevant and actionable!**

5. Conclusion

We started with a set of 78,000 rows and 19 variables of data extracted from our CRM and tried to intuitively interpret it with a managerial approach. We tried to understand what drives our sales and why we have not been converting enough deals.

We looked at some variables that could be strong indicators of our sales performance. We looked at sales results by sales amounts, region, deal size category and route to market. We gained some interesting insights but none of them uncovered success patterns. The best performance we could achieve was a modest 42% deals conversion with Telesales route to market and opportunities of less than 10,000 USD.

With such a large dataset, we couldn't realistically explore each and every variable to gain insights about what opportunities we can expect to win. We had to automatize our exploration process to determine the most significant features, which could strongly predict the opportunities results, specifically the won deals.

We tried Decision Tree and Random Forests models and achieved very good results. Random Forest yields an overall prediction accuracy of 88% and 77% accuracy on won deals. More importantly, Random Forest

could drastically reduce the dimension of our dataset and provide the most significant features for predicting the opportunities results. **Random Forest is a great fit for the job that we had in hands!**

So, we could interpret our large initial dataset in terms that our sales managers can understand.

We uncovered patterns about our opportunities, sales pipeline and what drives our win and losses. We built easy visualizations to help to **understand the profiles of the most likely successful sales opportunities**. For example, we realized that the chances of winning a deal decreases as it stays longer in the pipeline or that an opportunity is more likely to result in a win if the Client has purchased from us up to 50,000 USD.

These sales profiles are extremely valuable and more importantly, actionable in the hands of our sales teams. When reviewing their deal pipeline, our managers can anticipate gaps and correct their sales strategies accordingly. They can focus on the right deals and optimize their progression through the pipeline.

We can uncover more complex relationships by adding predictors according to their significance given by our predictive model, nevertheless we should always keep in mind the need of interpretability and at the end of the day, we want our **insights to be actionable by our sales managers!**

Note: In this project, we covered typical Data Science aspects with data wrangling (data collection, data tidying, feature engineering), data visualization, and machine learning.

Thank you for reading this report!