Capstone Project: Customer attrition for online retailer

piRsquare

12/17/2019

Executive summary

The main objective is to propose a prediction for customer attrition for a UK online retailer, using machine learning techniques. A secondary objective is to propose a linear regression to forecast next-month sales. This is proposed as secondary ojective given that it is not fully aligned with applying the machine learning techniques learned in class, although it is very relevant from a business perspective.

The original data is based on online transactions with a UK retailer, from Dec 2009 to Dec 2011. The data was published at the UCI repository by Dr. Daqing Chen (chend@lsbu.ac.uk), School of Engineering, London South Bank University. The data provides 1M records for invoice lines, for 54K invoices, 5K products, 6K customers.

For our main objective, we explore 3 different mathematical models, and 4 machine learning methods. Since we have different ways to formulate the problem predictors, we explore 3 mathematical models based on (1) monthly sales, (2) monthly activity, and (3) the Recency-Frequency-Monetary or RFM formulation. We apply 4 machine learning methods, namely, (1) KNN, (2) random forests, (3) logistic regression, and (4) Naive Bayes.

We randomly split 80% of the customers to train all 3X4 combinations, and to choose the best model/method. We reserve 20% of the records for testing final accuracy and as a way to verify the methods are not overfitted. The best combination is the logistic regression using the activity model with 71% accuracy during training, followed by the logistic with the RFM model. The sales model is outperformed, and hence discarded. The test set provides slightly different accuracy values, and confirms the methods are not overfitted.

For our secondary objective, we compare 2 forecasting methods based on linear regression. The forecasts are calculated at the customer level using a regression fitted to estimate the prior period sales, with all customers and their data for the prior year. The data for the period to be forecasted is not used to calculate the regression coefficients, and regressions are re-calculated when we move one period forward. One regression is based on using the data from all prior 12 months, and the second regression uses only the 3 months with the highest autocorrelation (i.e. the prior month, 11 months ago, 12 months ago).

We measure the performance by comparing the actual sum of sales vs the predicted, for Jan 2011-Nov 2011. In both methods, the predictions are off early in the year, but become better in the second half. We recommend the regression with 3 variables mostly on a business practical criteria. As expected, a regression with more variables has a better numerical fit, but it tends to consistently underestimate the actual sales in the second half of the year, where the peak sales occur.

Methods and Analysis

Preliminary counts

```
## # A tibble: 1 x 4
##   Number_records Number_invoices Number_products Number_customers
##            <int>           <int>           <int>            <int>
## 1        1067371           53628            5304             5943
```

Missing values

The data has 8 variables, and we start by finding if there are any missing data (NA) in the 5 variables we will need in the analysis; we exclude StcokCode, Description and Country which will not be used. We can see that we are missing a large number of Customer IDs, about 25% of the records. We cannot delete that many records, and later on we will show they represent an important part of the sales. Hence, we will treat them as if they were transactions with one generic anonymous customer identified with NA.

```
## # A tibble: 1 x 5
##   Invoice_NA InvoiceDate_NA Quantity_NA Price_NA CustomerID_NA
##        <int>          <int>       <int>    <int>         <int>
## 1          0              0           0        0        243007
```

Outliers for Price and Quantity

We will use Price and Quantity to calculate Sales value, hence we need to understand if outliers are data errors that need to be addressed. The table below shows the outliers for Quantity (i.e. 3 highest and 3 most negative). #1 seems a legitimate purchase; #3-4 and #7-8 are a order and its matching cancellation, with 0 net sales impact; the remaining #2, 5, 6 have Price = 0, we suspect they are inventory losses, with no impact on sales.

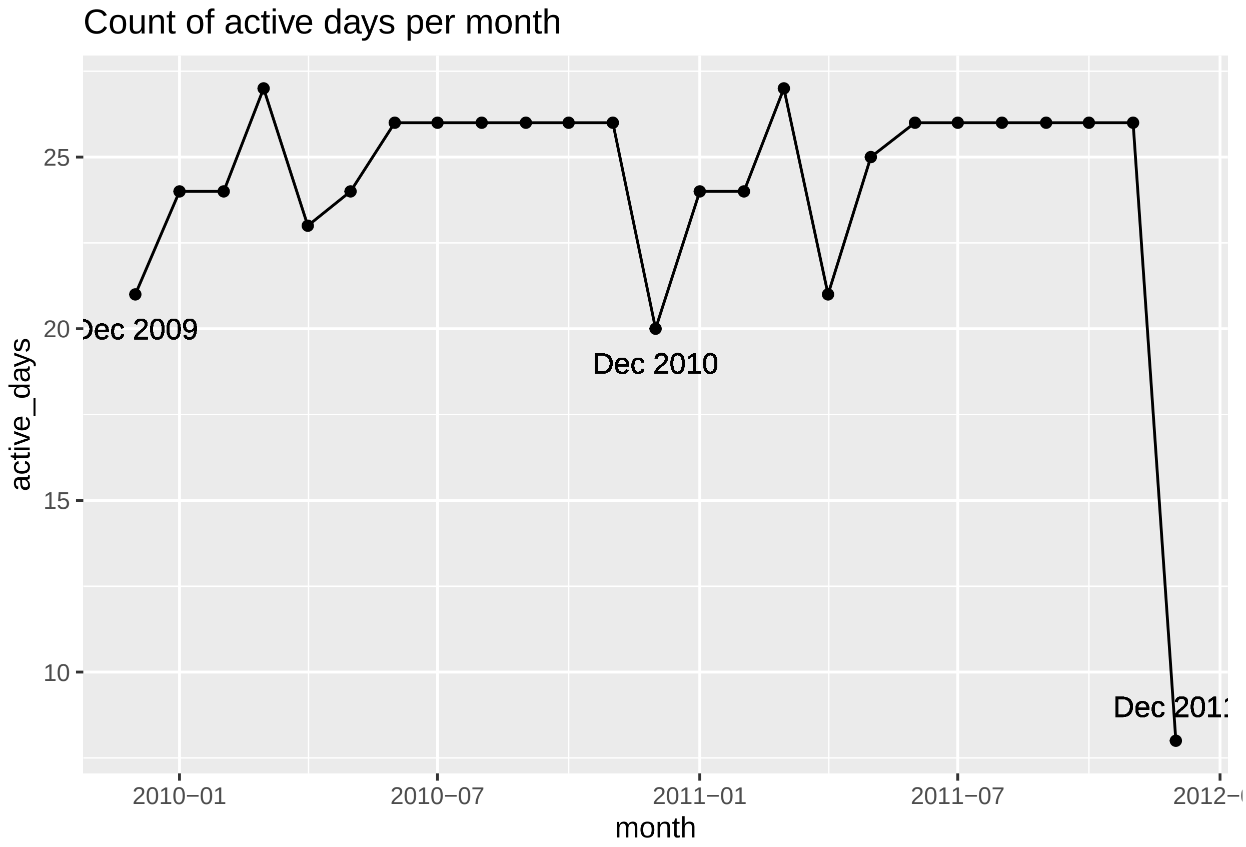
```
## # A tibble: 8 x 6
##   Invoice StockCode Quantity InvoiceDate         Price `Customer ID`
##   <chr>   <chr>        <dbl> <dttm>              <dbl>         <dbl>
## 1 497946  37410        19152 2010-02-15 11:57:00  0.1           13902
## 2 519017  22759        -9600 2010-08-13 09:14:00  0                NA
## 3 541431  23166        74215 2011-01-18 10:01:00  1.04          12346
## 4 C541433 23166       -74215 2011-01-18 10:17:00  1.04          12346
## 5 556690  23005        -9600 2011-06-14 10:37:00  0                NA
## 6 556691  23005        -9600 2011-06-14 10:37:00  0                NA
## 7 581483  23843        80995 2011-12-09 09:15:00  2.08          16446
## 8 C581484 23843       -80995 2011-12-09 09:27:00  2.08          16446
```

The next table repeats the outlier analysis using the Price variable. StockCode = B corresponds to adjustments for bad debt; StockCode = M are manual corrections. All normal part of business, hence we accept them.

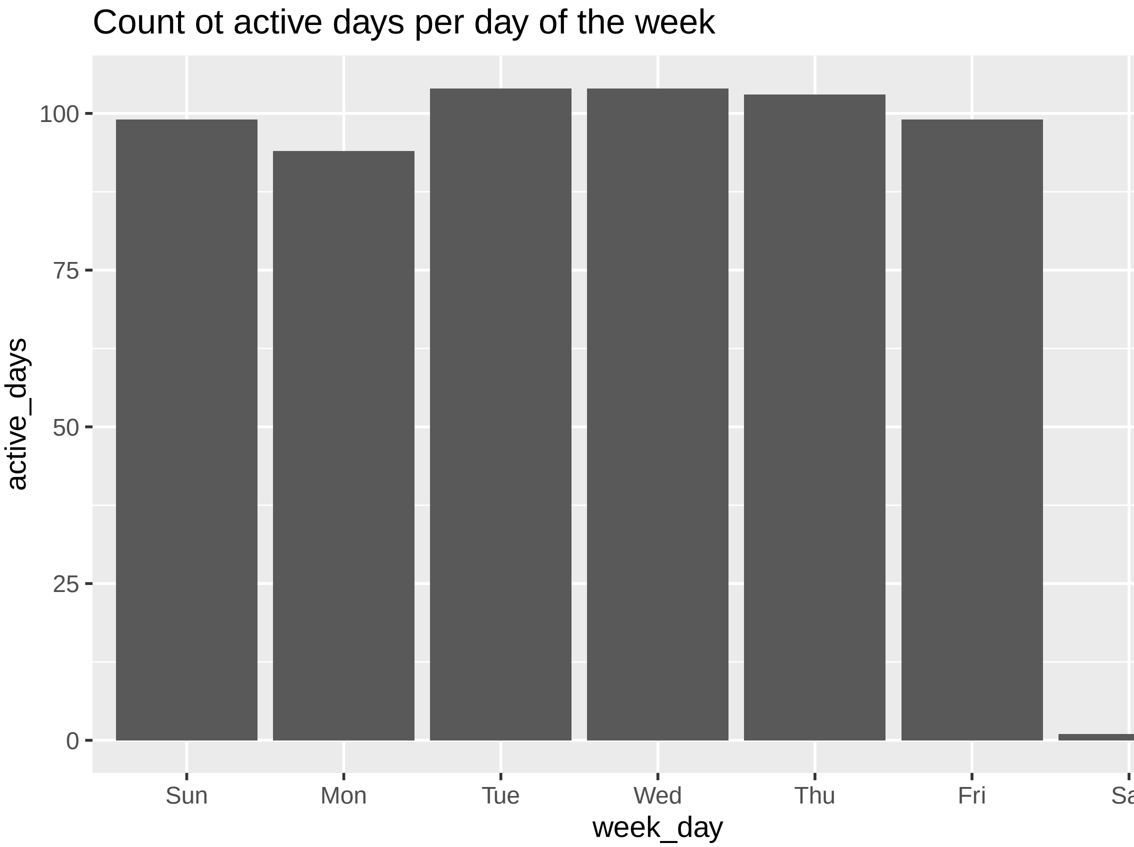
```
## # A tibble: 6 x 6
##   Invoice StockCode Quantity InvoiceDate           Price `Customer ID`
##   <chr>   <chr>        <dbl> <dttm>                <dbl>         <dbl>
## 1 A506401 B                1 2010-04-29 13:36:00 -53594.            NA
## 2 C512770 M               -1 2010-06-17 16:52:00  25111.         17399
## 3 512771  M                1 2010-06-17 16:53:00  25111.            NA
## 4 A516228 B                1 2010-07-19 11:24:00 -44032.            NA
## 5 A528059 B                1 2010-10-20 12:04:00 -38926.            NA
## 6 C556445 M               -1 2011-06-10 15:31:00  38970          15098
```

Available dates

We calculate the number of days with active sale per month, as a way to detect if we have incomplete months. Clearly, we have one month with incomplete data, Dec 2011, that we will have to exclude from the analysis. In 2009 and 2010, December has less active days because of the post-Christmas break.



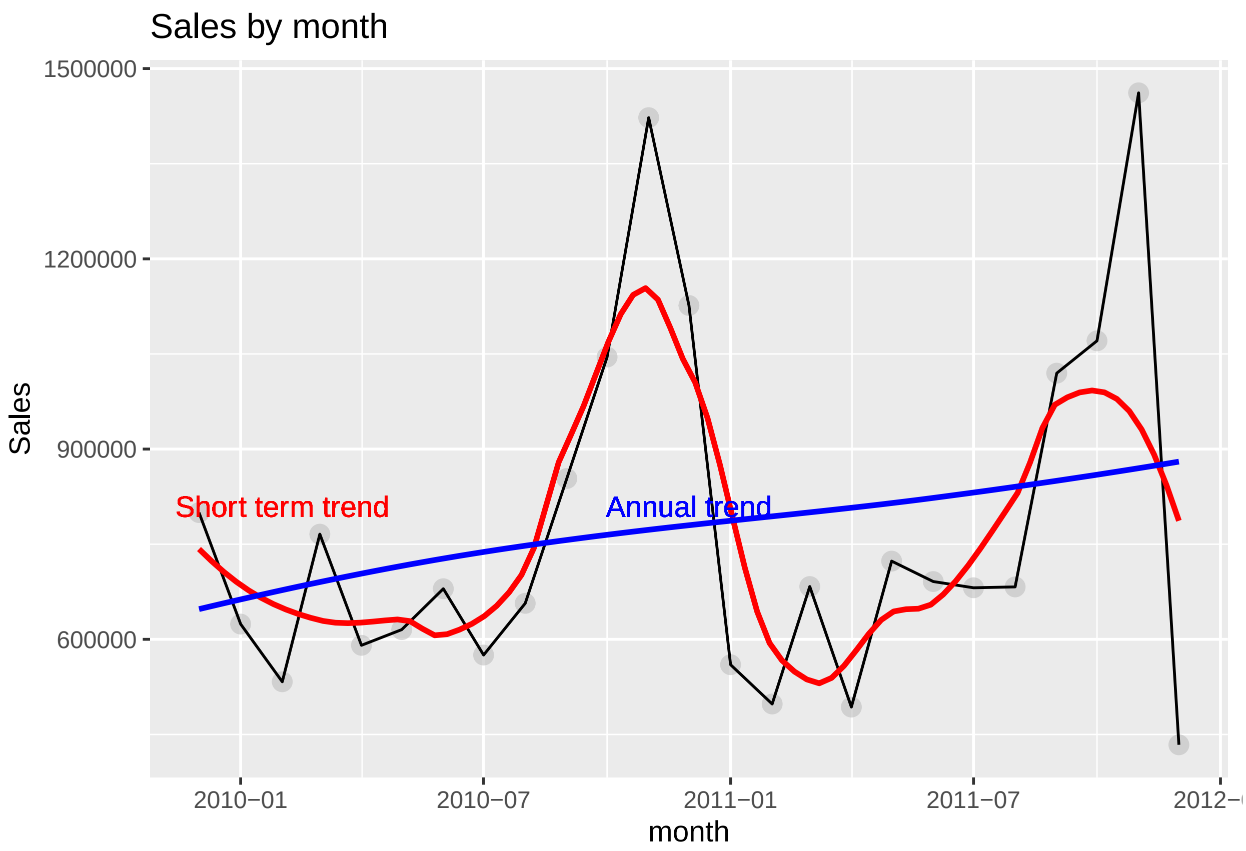
We also noticed that the company seems to take Saturdays off, and that explains why we do not see months



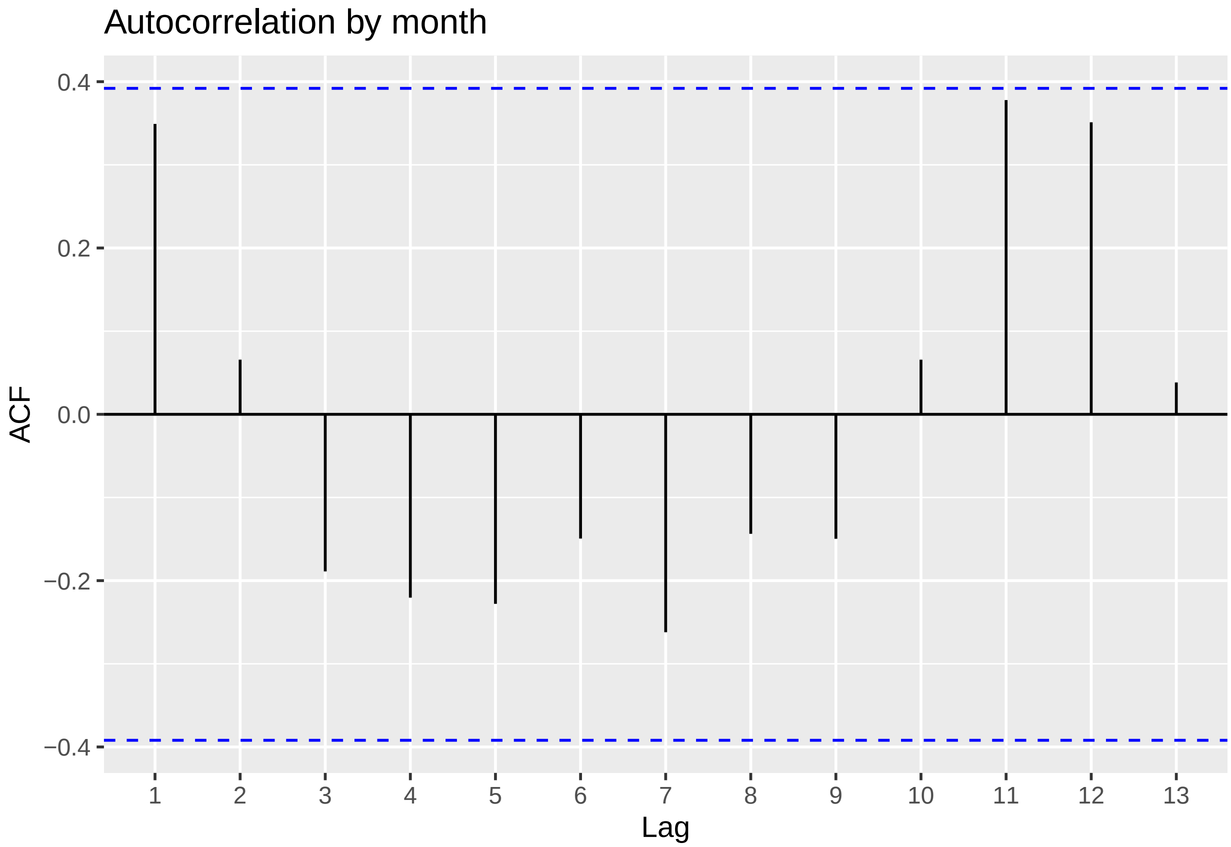
with 30 or 31 active days.

Sales seasonality

Sales have a strong seasonality, with a peak in the last quarter of the year. There seems to be a mild annual increase too. The seasonality implies the need to use models with a time horizon of 12 months.



We also calculated the autocorrelation of sales. Given the annual seasonality, it is no surprise to see a higher correlation with lags of 11 and 12 months. There is also some correlation with the prior month. This observation justifies to explore a forecasting model based on those 3 periods.



Pareto distribution of customer sales

A histogram of customer sales indicates there are a few large customers, and many small ones. In business, we usually hear that *20% of the customers represent 80% of the sales*, and we will show that that applies to oru UK online retailer too.

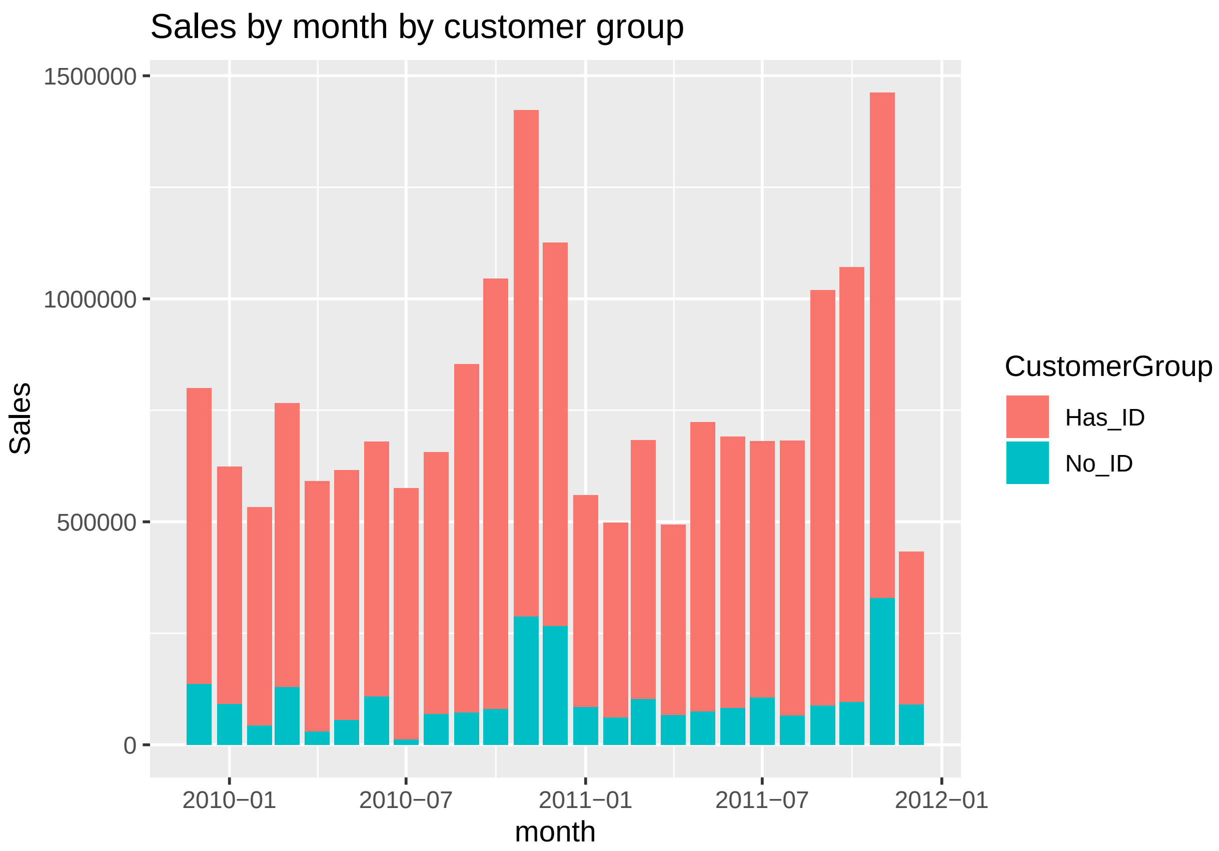
We can see that our top 5 customers are led by the group identified with NA, which is all customers missing ID.

```
top_n(customers,5)
```

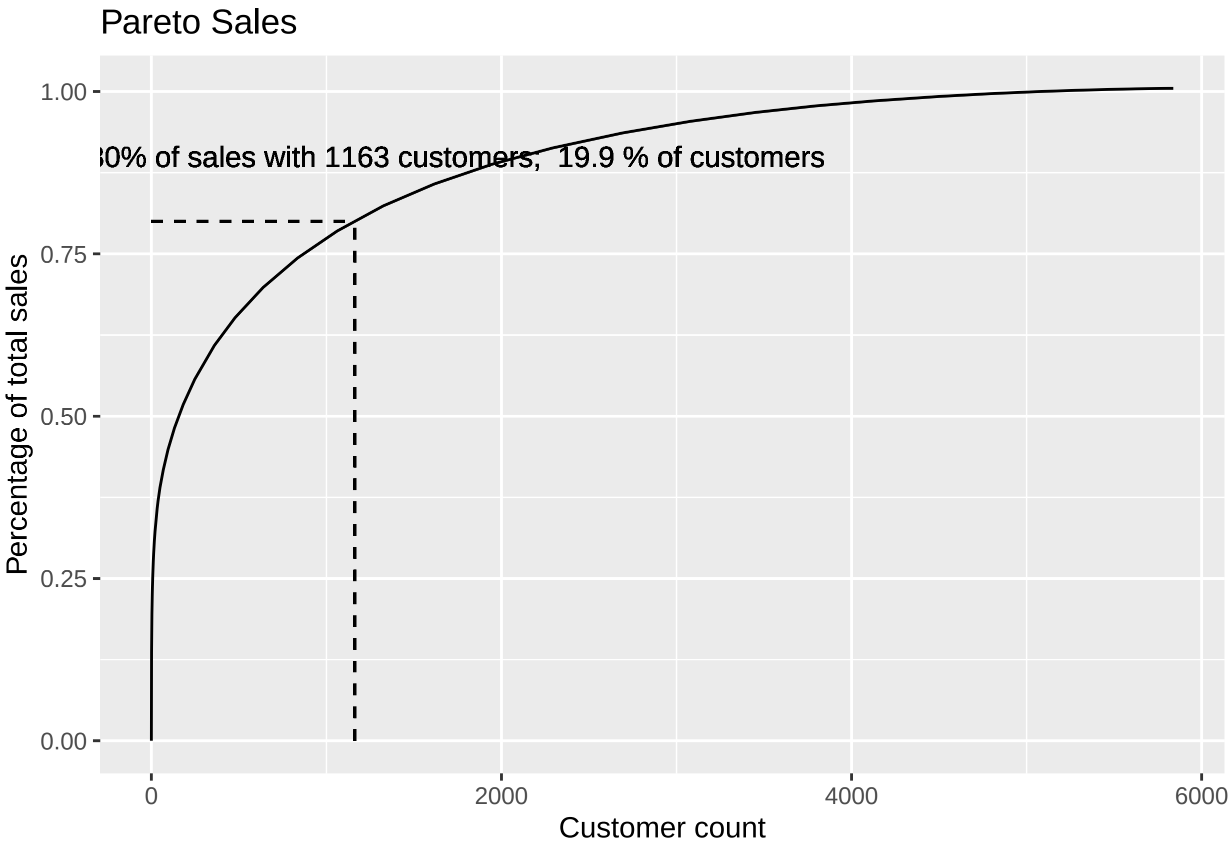
```
## Selecting by Sales
```

```
## # A tibble: 5 x 4
##   `Customer ID` LineItems    Qty    Sales
##           <dbl>     <int>  <dbl>    <dbl>
## 1            NA    243007 374364 2638958.
## 2         18102      1068 187110  598215.
## 3         14646      3890 365220  523342.
## 4         14156      4130 163910  296565.
## 5         14911     11613 143741  270249.
```

We can also see that sales to customers with missing ID are important and happen every month, making a case for keeping those records as part of the analysis.



We close our exploratory analysis with the Sales Pareto chart. When we sort the customers by sales, we confirm that the cumulative sales of the 20% larger customers represent 80% of the total.



Prediction of customer attrition

The main objective is to predict customer attrition:

- Define a customer to be attrited if 2011 Sales are zero or negative
- Use 2010 data to build three formulation *models*

  1. Sales, with 12 predictors defined as monthy sales in Jan-Dec
  2. Activity, with 12 predictors defined as 1 if Sales >-, 0 otherwise
  3. RFM, with 3 predictors

  - Recency = number of months between the last 2010 activity and Jan 2011, i.e, 1 if last positive sale was Dec 2010
  - Frequency = number of 2010 months with activity
  - Monetary = 2010 sales

- Split customers randomly into training and test sets, with 80% and 20% of 2010 customers, respectively
- Apply four machine learning *methods*

  1. KNN or K-Nearest Neighbors (cross-validated)
  2. Random forests (cross-validated)
  3. Logistic regression
  4. Naive Bayes

- Use training data to calculate method parameters and to evaluate prediction accuracy for the 12 model/method combinations
- Select model/method combination with highest accuracy

- Evaluate prediction accuracy in the test set, to check for overfitting

We provide an example in order to compare the 3 models

Model	Predictors	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Recency	Freq	Mon
Sales	12	0	100	0	0	150	-50	0	0	200	0	0	0			
Activity	12	0	1	0	0	1	0	0	0	1	0	0	0			
RFM	3													4	3	400

The highest accuracy is for the validation model with the logistic method, with RFM/logistic a close second. Detailed results will be discussed in the next section.

Prediction of next-month sales

A secondary objective was to explore linear regression to forecast next-month sales, i.e. to forecast 1-period ahead using past data. This constraint prevents from using the smootthing techniques learned in class, since they smooth a period using past and future periods.

- We use sales data by customer, by month
- Predictions are calculated by customer, and aggregated to find monthly sales
- Each month, a regression is trained to predict the current month using the prior 12 months, e.g. month 0 as a function of months -1, … , -12
- The periods are shifted one month, and the regression used to predict the next month, e.g. month 1 predicted using the regression for month 0 with the data for months 0, -1, … , -11
- Use two regressions, one based on all prior 12 months, a second based on the months -1, -11, -12 which have highest autocorrelation
- Calculate predictions for Jan-Nov 2011, to compare the models. Dec 2011 is a partial month data and cannot be used.

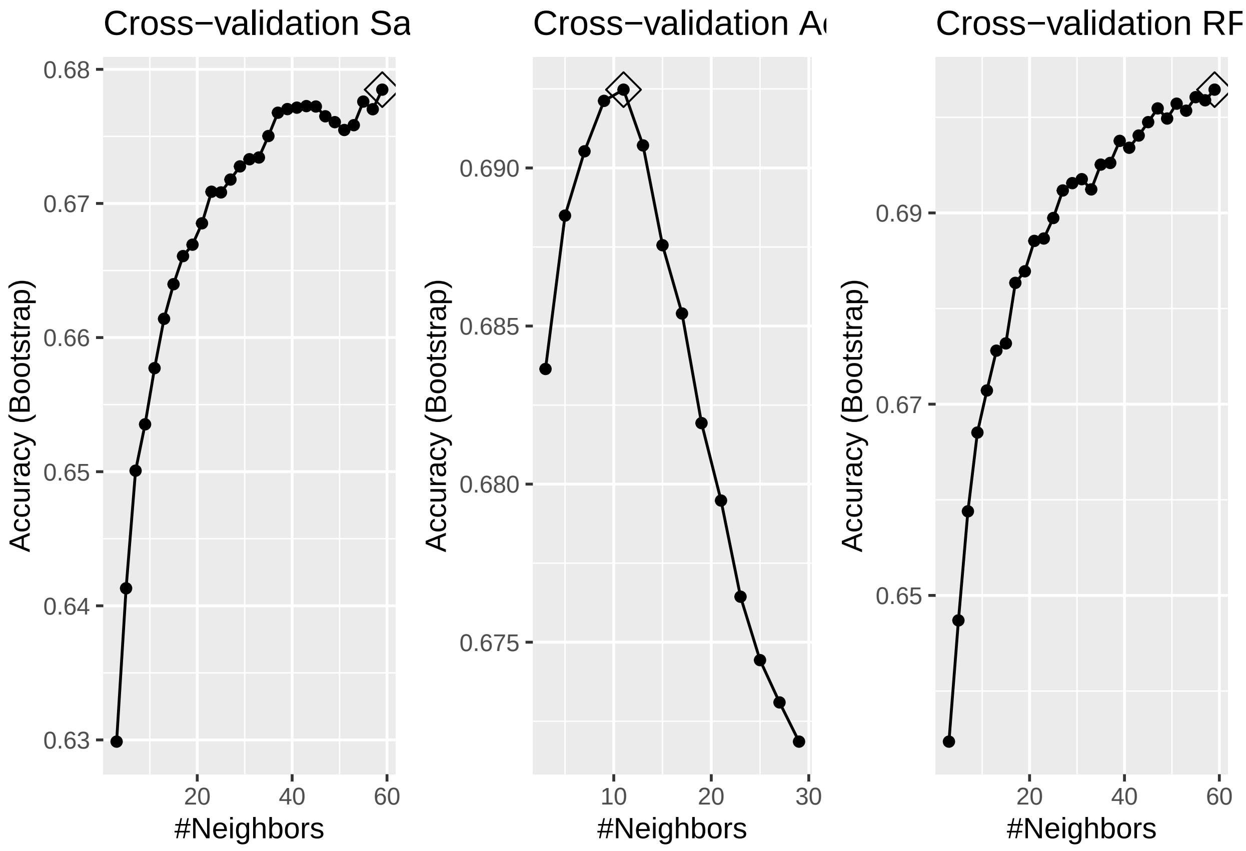
The approach departs from traditional machine learnign in that we do not split the data set randmly into a training and test set. Instead, we take advantage of the additional structure provided by the time variable. But we respect the principle that a model cannot be trained using the data it needs to predict. Results are discussed in the net section.

Results for Attrition Prediction

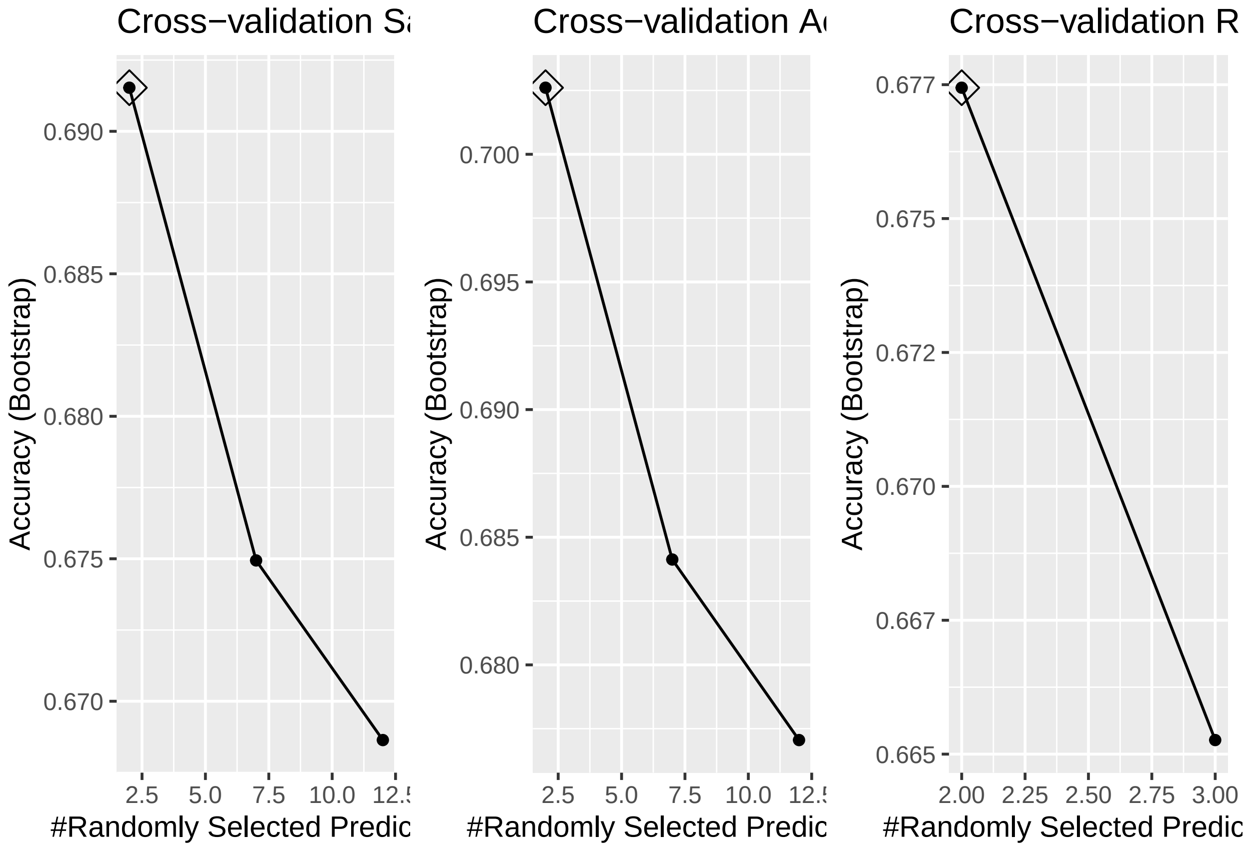
The following charts show the results of using cross-validation to select the best parameters for KNN and random forest.

- For KNN, we need to find the best k, we can see the activiy model prefers a smaller k, while the other 2 models prefer a larger k
- For random forest, all the best for all models is to use 2 random predictors

Cross-Validations KNN method



Cross-validation Random Forest method



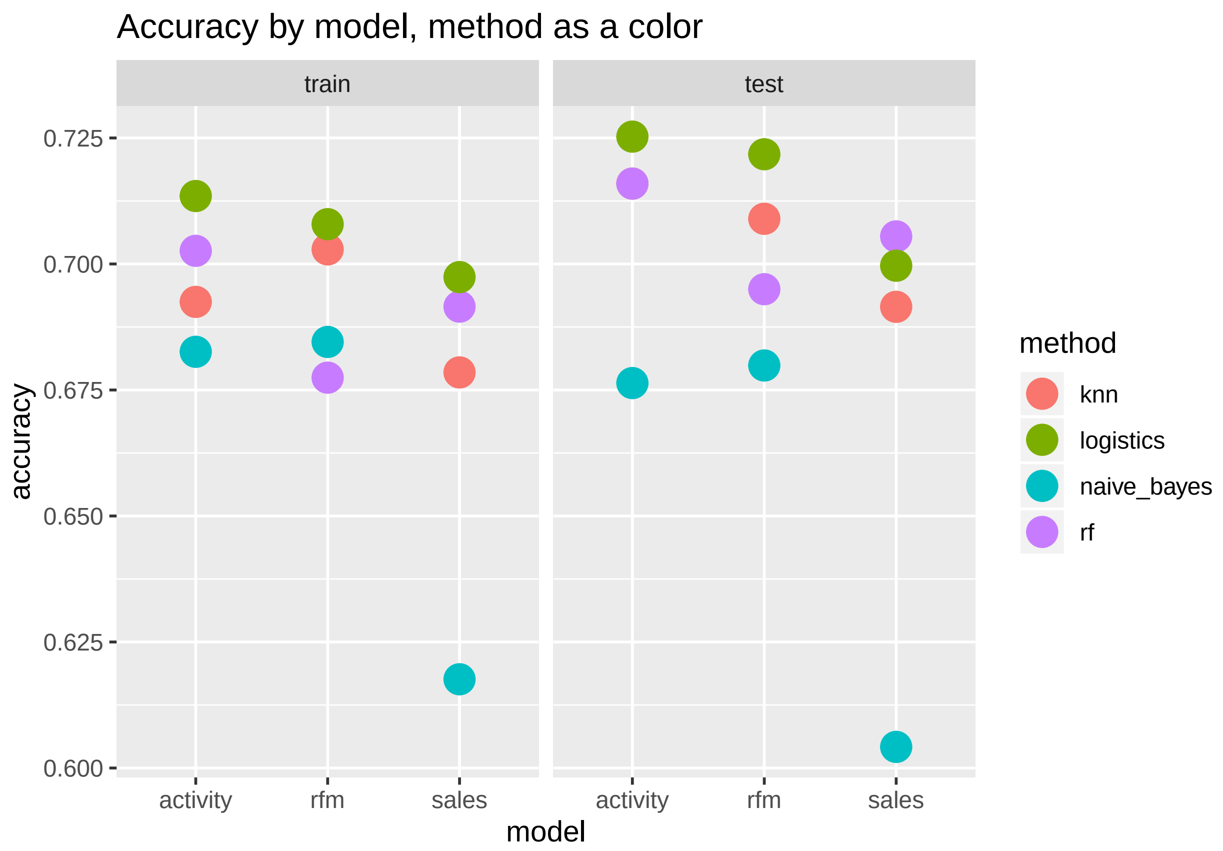
Final Results for Attrition Prediction

For each of the 3 models, we have accuracy results of 4 learning machine methods. A simple average by model seems to indicate the Sales model is the weakest.

```
##    sales activity      rfm
##    0.671    0.698    0.693
```

The table above is only an indication, but more detailed results confirm the relative weakness of the sales model. The charts below provide more details.

- The best model is chosen using the training data on the left chart, and it is the logistic with the activity model with accuracy 71.3%
- The second best model is the logistic with the RFM model with accuracy 70.8%
- The accuracy on the test set confirms there is no overfitting. Note that kkn and rf have the same test accuracy for activity, the dots overlap and only one shows.

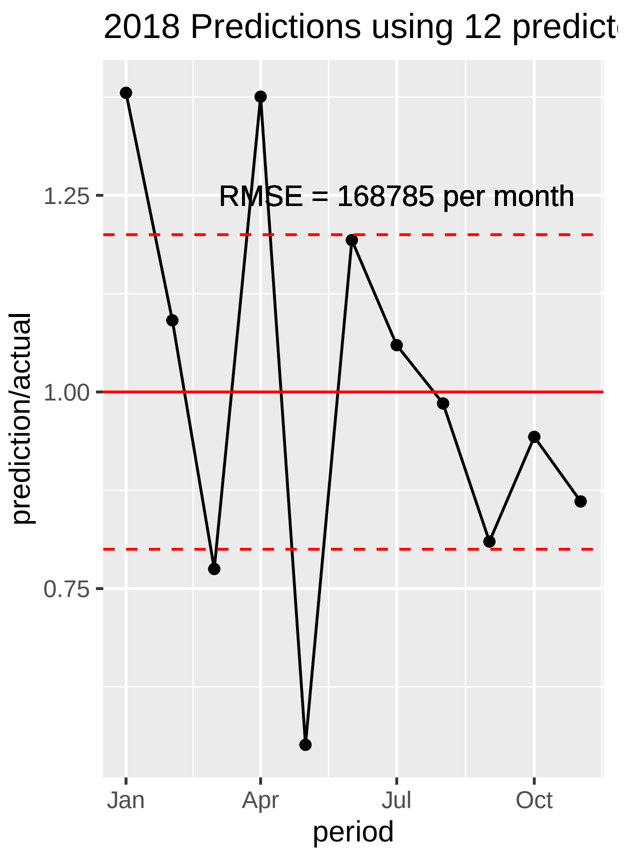


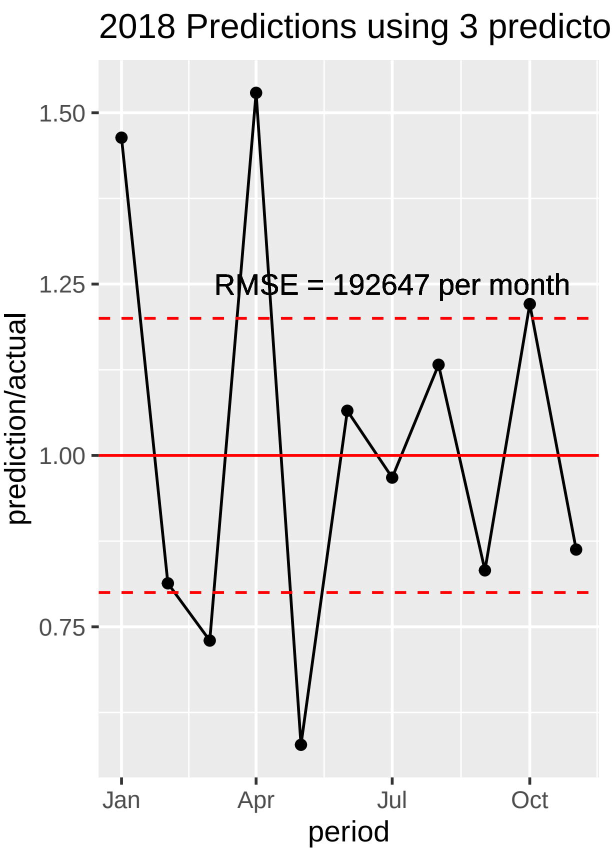
Final Results for Next-Month Sales Forecast

The next-month forecast was our secondary objective. We understand it is not fully aligned with the materials studied, but we think it has business relevance. The charts show the results of comparing prediction with actuals for 2011. As explained before, we compare a regression with 12 months vs a regression with 3 months. Graphically we can see that:

- Both Forecasts can be off by more than 20% in some months
- Forecasts improve in the second half of the year, which is good news because it is when the sales peak
- The regression with 12 months have bias to underperform in the second half of the year

Based on those observations, we recommend to use the forecast with 3 months, with the caveat that we should continue searching for a more accurate approach.





Conclusions

We covered two objectives, to define a prediction model for customer attrition using machine learning, and to explore a next-month forecast for sales, for a UK online retailer.

The main learning from this project was about data and Rmd wrangling. The majority of time and effort was in finding the right command or option to make the code work. Looking for answers in the help files, online tutorials, and comments from the global community was a great way to expand my knowledge of R and R markdown.

The second learning is that it is important to pick the right formulation for the variables in your problem. In this case, transforming the Sales variable into an Activity variable helped improve the attrition prediction accuracy. Please note that attrition is a binary variable, so is activity. That might explain why sales, a continuous variable, underperforms. RFM also does well to predict attrition, it is a mixed model with 2 binary variables and one continuous (monetary).

The third learning is that there is too much to learn whether there is a conection between a data and its structure that could help identify the best machine learning technique to use. There is an exhuberancy of machine learning algorithms, and the approach is to try many to see which one fits best. Is there a better way?

In terms of potential future enhancements, the attrition model could be improved if there was a way to clean up the missing customer IDs. It also would be interestiting to look for data beyond 2011 to validate and test.

The next-month sales forecast might be improved with additional years of data. Seasonality could be made explicit with a multiplicatitive model, but more years of data are needed for a robust calculation.