**CAPSTONE PROJECT**

**Analysis of Customer Churn prediction in Logistic Industry**

**MENTOR: TEAM MEMBERS:**

ANKUSH BANSAL UTKARSH NIPANE

DATA SCIENTIST SIDDHARTH KHANNA

SHIVAM VERNEKAR

MANOJ KAUSHIK

**CONTENTS**

[**CHAPTER 1** 3](#_Toc8672)

[**INTRODUCTION** 3](#_Toc8673)

[**1.1** **CUSTOMER CHURN ANALYTICS** 3](#_Toc8674)

[**1.2** **SCOPE OF THE PROJECT** 3](#_Toc8675)

[**1.3** **VARIABLES CONSIDERED FOR ANALYSIS** 4](#_Toc8676)

[**CHAPTER 2** 5](#_Toc8677)

[**DATA CLEANING** 6](#_Toc8678)

[**2.1 DATA SET CHOSEN** 6](#_Toc8679)

**2.2 MISSING VALUE TREATMENT**………………………………………………………………………………………………….6

**2.3 OUTLIER TREATMENT**……………………………………………………………………………………………………………..6

**2.4 FEATURE ENGINEERING**……………………………………………………………………………………………………….….7

**2.5 SEGMENTATION USING UNSUPERVISED LEARNING …………………………………………. 8**

[**EXPLORATORY DATA ANALYSIS** 8](#_Toc8681)

[**3.1 INTRODUCTION** 8](#_Toc8682)

**3.2 EDA USING TABLEAU** 8

**3.3 INFERENCES WITH RESPECT TO THE TARGET VARIABLE:** 12

[**CHAPTER 4** 13](#_Toc8680)

**4.1** [**BASE MODEL FOR PREDICTING CUSTOMER CHURN** 13](#_Toc8681)

[**4.1.1 LOGISTIC REGRESSION MODEL** 13](#_Toc8679)

[**4.1.2 DECISION TREE MODEL** 15](#_Toc8679)

**4.2 CHOOSING THE MACHINE LEARNING CLASSIFIER**………………………..…………………………………………..16

[**REFERENCES** 21](#_Toc8692)

**CHAPTER 1**

# **INTRODUCTION**

## CUSTOMER CHURN ANALYTICS

Customer churn prediction in logistics industry is one of the most prominent research topics in recent years. It consists of detecting customers who are likely to cancel a subscription to a service. Recently, logistics market has changed from a rapidly growing market into a state of saturation and fierce competition. The focus of the logistic companies has therefore shifted from building a large customer base into keeping customers in house. For that reason, it is valuable to know which customers are likely to switch to a competitor in the near future. The data extracted from the industry can help analyse the reasons of customer churn and use that information to retain the customers. We have proposed to build a model for churn prediction for a company using data mining and machine learning techniques namely logistic regression and decision trees. A comparison is made based on efficiency of these algorithms on the available dataset.

## 1.2 SCOPE OF THE PROJECT

Churn rate is the number of customers or subscribers who cut ties with your service or company during a given period. These customers have “churned.” The scope of the project is to use RFM to classify customers and do factor engineering and build a model that can predict churn. The method can be used for campaign management, by modelling the best groups to be approached with a specific marketing campaign and by pinpointing individuals who are the most influential over their peers. Our method can also be used in other domains where the links between people can be measured or inferred, such as social networking sites on the Internet, in order to predict customer behaviour.

# **1.3 VARIABLES CONSIDERED FOR ANALYSIS**

The following variables are considered for the analysis:

* **Customer ID**: The five digits unique identifier representing a customer.
* **LastInvoiceDate**: The format is DD-MM-YYYY. The date on which an invoice for a good is issued, which is usually the same day the good is sent to the buyer.
* **AvgInvoiceAmount**: An average amount of invoice, bill or tab, which is a commercial document, issued by a seller to a buyer, relating to a sale transaction and indicating the products, quantities, and agreed prices for products or services the seller had provided the customer.
* **LocationName:** The name of a geographical [location](https://www.collinsdictionary.com/dictionary/english/location), such as states and city, the first term is name/abbreviation of the city and term after punctuation is the name of the state of the region customer belongs.
* **Zip:** a postal code consisting of five digits of the region customer belongs.
* **AvgMiles:** The distance on the average travelled by a Logistics service provider.
* **NoCalls:** Number of Calls being made by the customer regarding any queries and complaints.
* **ClaimsMade:** Number of claims made which are the policies providing coverage that are triggered when a claim is made against the insured during the policy period, regardless of when the wrongful act that gave rise to the claim took place
* **DelayedQuote:** A stock or other quotes that are reported some time after the transaction takes place where 0 represents “No Delayed Quote” and 1 represents “Delayed Quote”.
* **PickupDelay:** Number of delays happened during pickups where 0 represents “No Pickup Delay” and 1 represents “Pickup Delay”.
* **DeliveryDelay**: Number of shipments arrived later than the time that they were scheduled to arrive where 0 represents “No Delay” and 1 represents “Delay”.
* **VolumeChange**: The tendency of matter to change in volumewith response to a change in temperature where 0 represents “No Change” and 1 represents “Some Change”.
* **PriceChangedPostQuote**: A quote (or quotation) is an exact price for the job being offered. As such it is fixed and cannot be changed once it has been accepted by the customer unless the customer changes the amount/type of work required or you discover something completely outside of the scope of what was agreed where 0 represents “No Price Change” and 1 represents “Price Changed post Quote”.
* **CompetitorsPresent**: Whether there is any person or entity which is a rival against another where 0 represents “No Competitors Present” and 1 represents “Competitors are Present”.
* **CurrentInflation**: Whether there is an increase in the cost of living as the price of goods and services rise, where 0 represents “No Current Inflation” and 1 represents “Current Inflation has happened”.
* **Customer Churn**: This variable indicates whether the customers have “churned” or not. The victory the Supreme Court provided the petitioning party may not have been total and complete. **Overall percentage of customer churn** in given dataset is **10.8%.**

**LIMITATIONS WITH THE DATASET:**

* The dataset is heavily imbalanced with respect to the target variable.
* The number of repeat customers is very insignificant and thus affects the RFM score considerably.

**TOOLS AND TECHNIQUES:**

* **Jupyter Notebook** for coding
* **Tableau** for Data Visualization

**CHAPTER 2**

**DATA CLEANING**

**Problem Statement**

Using the data set, customer\_data.csv of a premier Shipping and Logistics company, use RFM to classify customers and do factor engineering and build a model that can predict churn.

**2.1 DATA SET**

The dataset provided is about Customer Churn. It consists of 600 records and 16 features. Provides insight into the details of the customers including Volume Change, Competitors Present, Current Inflation, Claims made etc. “Customer Churn” attribute conveys whether the customer has churned or not.

Customer churn analysis refers to the customer attrition rate in a company. This analysis helps companies identify the cause of the churn and implement effective strategies for retention.

**2.2 MISSING VALUE TREATMENT**

There are no missing values present in the dataset.

**2.3 OUTLIER TREATMENT**

There are outliers present in” AvgMiles” and” AvgInvoiceAmount” features.

There were 7 outliers present in “AvgInvoiceAmount“ and 6 outliers present in ”AvgMiles” columns which is not that significant and thus have been treated byusing mean value imputation.

**2.4 FEATURE ENGINEERING**

Feature engineering is the process of using domain knowledge of the data to create features that make machine-learning algorithms work.

## The feature “LocationName” contains the name of the city and the state. Using feature engineering, the states have been separated from the cities as a separate column for a much broader and better perspective. Thus, two new columns i.e. “state” and “city” are generated.

The feature **“Region”** is also included in the dataset to accommodate representative regions as **West, South, Northeast** and **Midwest** in United States of America.

**RFM (recency, frequency, monetary)** analysis is a marketing technique used to determine quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary).

* The recency column has been created with the help of “LastInvoiceDate” feature, which describes when the customers purchased.
* The frequency column has been created based on the repetitions of “Customer ID” which conveys how often the customers have purchased.
* The monetary values have been taken into consideration using the “AvgInvoiceAmount” which denotes the amount spent by the customers on an average.

**2.4.1 IMPORTANT FEATURES:**

* While doing the model building the most important features to stand out were **“ClaimsMade”, “VolumeChange”, “PriceChangedPostQuote” and “DelayedQuote”** according todata interpretation.

**2.5 SEGMENTATION USING UNSUPERVISED LEARNING:**

**Technique Used: K-Means Clustering**

RFM segmentation works really well to come up with a method to reward the customers something which does not involve any sort of prediction but rather action based on what they have done in a given time interval. However, when this same system is used to predict their behaviour in the next quarter, which is when it becomes a tricky scenario. For starters, this kind of segmentation does not capture what motivations or preferences these customers have.

In such a scenario, RFM is not the best way forward thus; we are using K-Means clustering to segment the data into two clusters with customer churn as 0 and non-customer churn as 1.

I.e. Customer not churned = 542 and customer churn = 58

Thus, K-Means clustering is properly able to segment the data into two above-mentioned categories.

**CHAPTER 3**

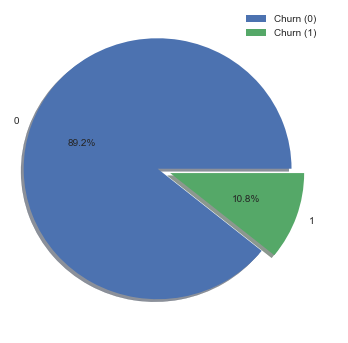
**EXPLORATORY DATA ANALYSIS**

**3.1 INTRODUCTION:**

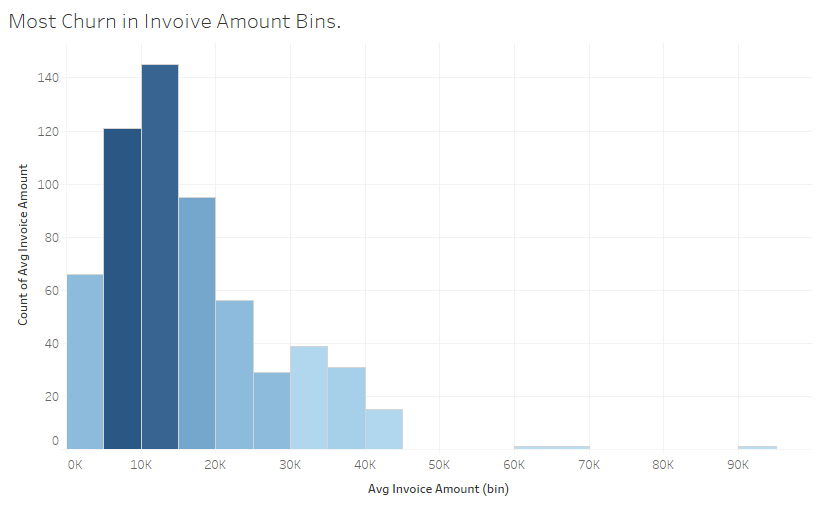
EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data.

**3.2 EDA USING TABLEAU:**

**3.2.1 Customer Churn Overall Percentage:**



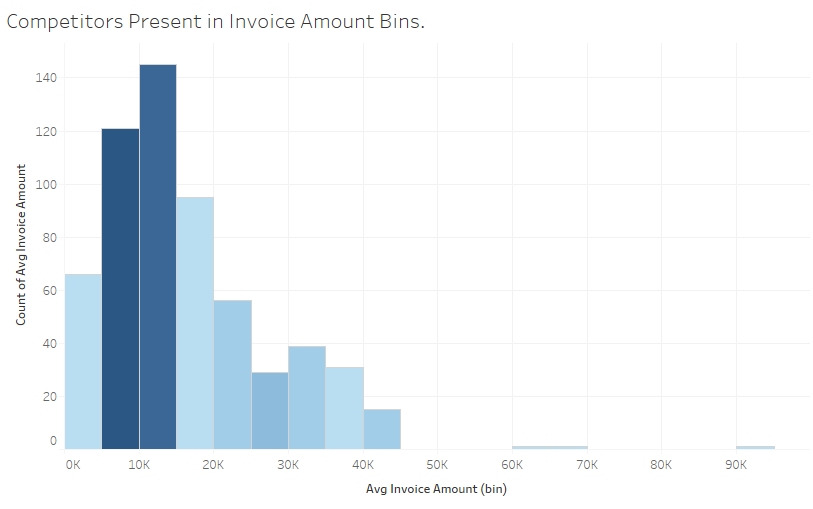
**3.2.1 Customer Churn in Invoice Amount Bins:**

****

Most of the customer churn is happening in the **Average invoice Amount range of 5K to 15K**.

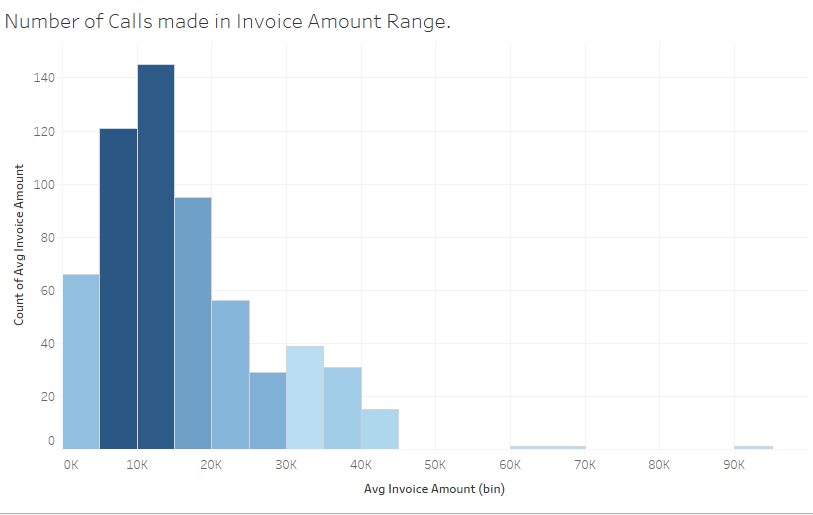
45 customers are churning out in this range out of the 65 customers which have churned i.e**. ~ 70% of the customer churn** is happening in this range itself.

**3.2.2 Competitors Present in Invoice Amount Bins.**



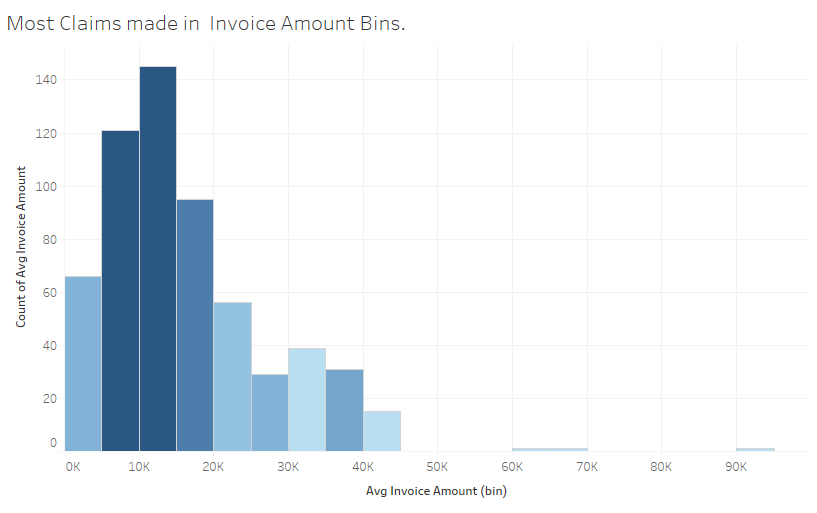
Most competitors are present in the **Invoice amount range of 5K to 15K**. 15 out of the total 20 competitors are present in this range i.e. **75% competitors are present** in this range.

**3.2.3 Number of calls made in Invoice Amount Bins.**



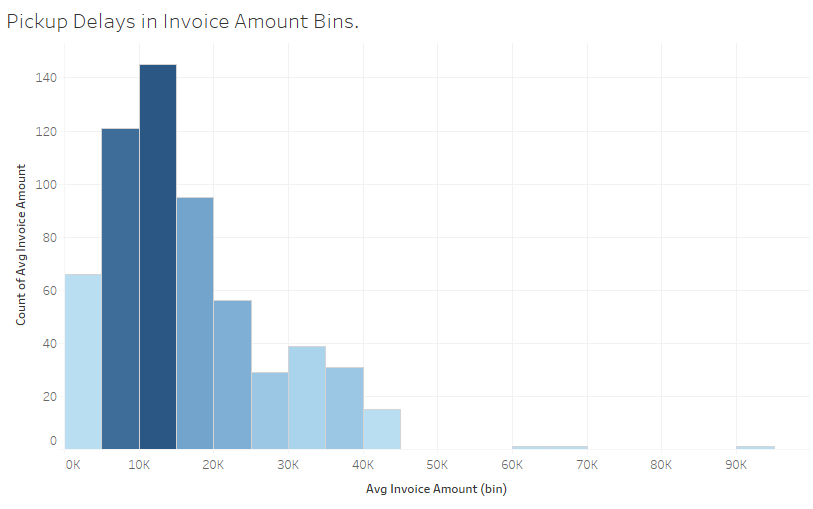
Number of calls made by the customers are high in **the Invoice Amount Range of 5K to 15K**. 97 out of the 169 calls made have been made in this range. i.e. **~ 58% of the calls** have been made on this range. The calls made by a customer either would be for complaints or for the inconvenience caused to the customer.

**3.2.4 Most Claims Made in Invoice Amount Bins**.

****

27 out of 34 claims have been made by the customers in the **range of 5K to 15k** i.e. **~80% of the claims** are made in this range.

**3.2.5 Pickup Delay in Invoice Amount Bins.**

**.**

28 out of 37 Pickup Delays have been found in the **range of 5K to 15K**. i.e. **~ 76% of the pickup delays** are found in this range.

**Conclusion from above Bar Graphs:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Percentage of variables in Avg Invoice Amount**  **5K to 15K (bins)** | **Actual Number**  **In given bin percentage** | **Total Number** |
| **Customer Churn** | **70%** | **45 Customers** | **65 Customers** |
| **Percentage of competitors are present** | **75%** | **15 competitors** | **20 competitors** |
| **Percentage number of calls made** | **58%** | **97 calls** | **169 calls** |
| **Percentage of Claims** | **80%** | **27 claims** | **34 claims** |
| **Pickup Delays** | **76%** | **28 Pickup Delays** | **37 Pickup Delays** |

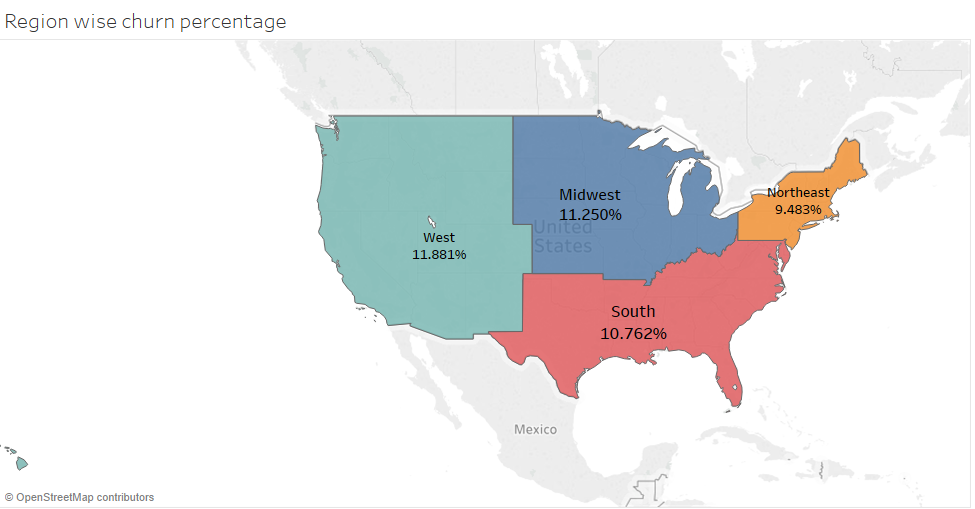
#### **3.2.5 Pickup Delay in Invoice Amount Bins**

#### 

#### We see that the state of **California and New York** have the high number of customer churn.

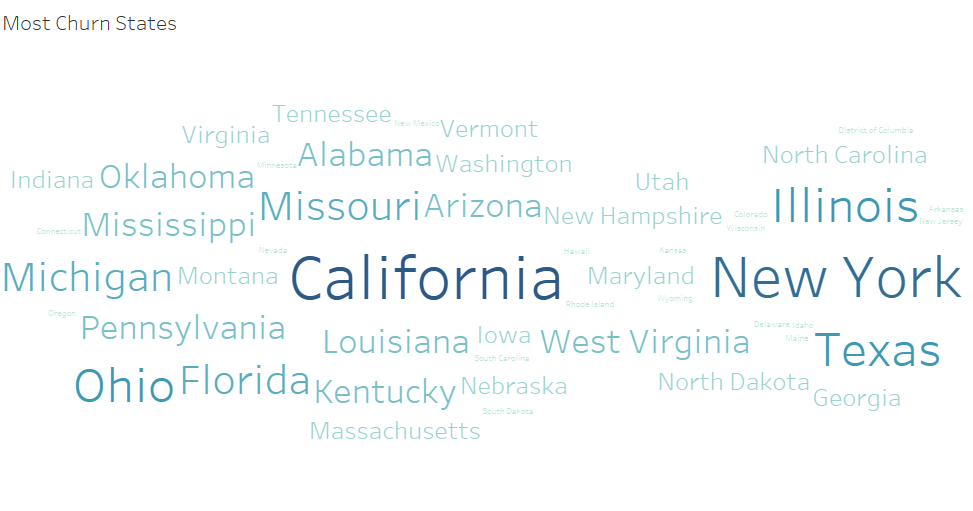
Geographic Distribution Of Customer Churn.

**3.2.6 Region-wise Churn Percentage: Highest Churn in West Region (11.88%)**



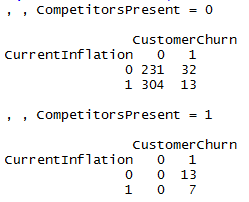
We see that the churn percentage in the four regions is the similar to the overall churn in the country. The highest percentage of Customer Churn is seen in the Western Region.

**3.2.7 Highest Churn States: California and New York**



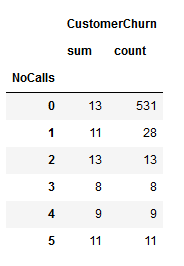
The cities in which we see the most churn are highlighted in the word cloud. Among the states California, New York, Texas etc. have high churn.

#### **3.3 INFERENCES MADE ABOUT THE FEATURES WITH RESPECT TO THE TARGET VARIABLE:**



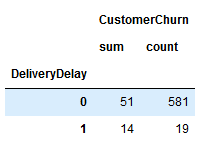
When we observe the cross table, we see if competitors are present, there is a 100% chance of customers getting churned as per the data.

* The most number of calls made by a customer is at most 5. We see 11 customers who have made 5 calls.

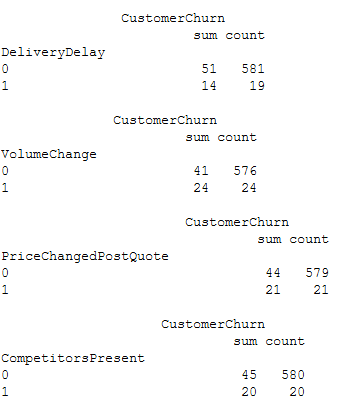


**As we see, if the Number of calls made by a customer is more than 1, there is a 100% chance of a customer getting churned.**

* We see that the customers who have made 2 or more than 2 calls have resulted in churn. More the number of calls the probability of churn is greater.
* The customers whose product delivery was late have resulted in churn. The probability of churn is greater if the delivery was late.



* **When there is a delivery delay, we see a lift in the probability of a customer being churned.**
* The customers who have made claims even once have resulted in churn. The following factors are responsible for churn:



For all the features present we can see the lift in probabilities of various features present.

* The customers who is product's Volume changed from ordering to delivery have resulted in churn. Although we see that out of all the customers who have resulted in churn, we see that the number of customers who is product volume did not change are more in number.
* Out of the customers who have resulted in churn. There are more number of customers whose pickup was not delayed. We see that if the Pickup is delayed and Delivery is delayed the customer churn is greater.
* The number of customers resulting in churn is more in number where the prices did not change post Quote. If the prices have changed, the customers have resulted in churn.
* If there are no competitors present in absence of inflation, the customer churn is higher than in the presence of inflation and competitors.

**CHAPTER 4**

### 4.1 Base model for predicting the Customer Churn in our data.

**4.4.1 FEATURE SELECTION:**

* The recency, frequency and monetary variables have been created for clustering and segmentation to find out the customers and their respective RFM scores. They are not being used for model building.

## The “LastInvoiceDate” has been further classified into Days, Months and Year. This has been to better the comprehend and help provide much better insights for the frequency and recency columns that have been created so, these feature were not significantly contributing in the model hence not used.

| * The percentage of Customer Churn with regards to the target variable is around 11 %. So, Features implemented for the model are “**AvgInvoiceAmount”, “AvgMiles”, “NoCalls”, “ClaimsMade”, “DelayedQuote”, “PickupDelay”, “DeliveryDelay”, “VolumeChange”, “CompetitorsPresent” and “CurrentInflation”.** |  |
| --- | --- |

**4.1.2 LOGISTIC RERESSION MODEL**

We built a Logistic Regression model to predict if a customer will churn or not.

Below were the results we got:

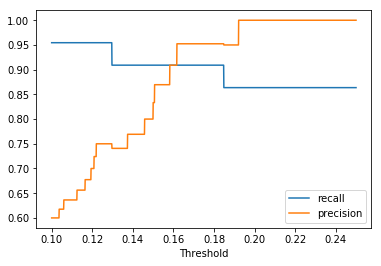
**Precision recall f1-score support**

**0 0.96 1.00 0.98 158**

**1 1.00 0.73 0.84 22**

* We see a high f1-score for 0 but at the same time f1-score is very low for 1. This is because of the high-class imbalance we have in our data.
* This can be dealt by either producing synthetic data using the original data or by changing the probability threshold values.

We tried to check the precision and recall score by changing the probability threshold values and got the following results:



# From the above plot, we can see Recall and precision values for different probability thresholds.

* The best threshold point **is around 0.17** where precision and recall are both high.
* So after using probability threshold as .17, we get the below **classification report:**

**Precision recall f1-score support**

**0 0.99 0.99 0.99 158**

**1 0.95 0.91 0.93 22**

* Here we can see a significant raise in the f1-score for both 0 and 1 by changing the thresholds.

**4.1.2 DECISION TREE MODEL:**

We also built a decision tree model to predict customer churn where choosing max\_depth=4, the model achieved better accuracy than Logistic Regression and got the below classification report:

**For validation data:**

**Precision recall f1-score support**

**0 0.99 1.00 0.99 158**

**1 1.00 0.91 0.95 22**

**For training data:**

**Precision recall f1-score support**

**0 0.99 1.00 1.00 377**

**1 1.00 0.93 0.96 43**

When comparing results of Decision tree and Logistic Regression, Decision tree turns out to be the winner.

**Decision tree has certain advantages over others:**

* Decision trees implicitly perform variable screening or feature selection
* Decision trees require relatively little effort from users for data preparation
* Variable transformations are not required with decision trees because the tree structure will remain the same with or without the transformation.
* Decision trees are very intuitive and easy to explain

# **4.2 Choosing a Machine Learning Classifier**

**How do you know what machine-learning algorithm to choose for your classification problem?**

* We really care about accuracy, so our best bet is to test out a different machine learning models (with different parameters within each algorithm), and select the best one by cross-validation.
* But here are some general guidelines we found to work:
  + If the training set is small, high bias/low variance classifiers (e.g., Naive Bayes) have an advantage over low bias/high variance classifiers (e.g., kNN), since the latter will over fit.
  + However, low bias/high variance classifiers start to win out as training set grows (they have lower asymptotic error), since high bias classifiers are not powerful enough to provide accurate models.

**4.3 Comparing different machine learning models to predict customer churn:**

We will try to check with the following classification models and check for the one with maximum accuracy:

1. **Logistic Regression**

**Advantages of Logistic Regression:**

* Many ways to regularize the model and you do not have to worry as much about your features being correlated, as if you do in Naive Bayes.
* There is probabilistic interpretation, unlike decision trees or SVMs, and you can easily update your model to take in new data unlike decision trees or SVMs.
* If required we can use a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you’re unsure, or to get confidence intervals) or if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model.

1. **Decision tree**

**Advantages of Decision Tree:**

* Decision trees implicitly perform variable screening or feature selection
* Decision trees require relatively little effort from users for data preparation
* Variable transformations are not required with decision trees because the tree structure will remain the same with or without the transformation.
* Decision trees are very intuitive and easy to explain

1. **Naïve Bayes**

**Advantages of Naïve Bayes:**

* The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Super simple, you’re just doing a bunch of counts.
* If the NB conditional independence assumption actually holds, a Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data.
* Even if the NB assumption doesn’t hold, a NB classifier still often does a great job in practice.
* A good bet if want something fast and easy that performs pretty well.

1. **Random Forest**

**Advantages of Random Forest**

* This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts.
* One of benefits of Random forest, which excites me most, is the power of handle large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods.
* Further, the model outputs Importance of variable, which can be a very handy feature (on some random data set).
* It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
* It has methods for balancing errors in data sets where classes are imbalanced.
* The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.

**After applying these 4 models and using K-Fold Cross Validation we got the following mean accuracy scores and standard deviation**.

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Accuracy** | **Standard Deviation** |
| **Logistic Regression** | 0.968333 | (0.023214) |
| **Random Forest** | 0.976667 | (0.022608) |
| **Decision Tree** | 0.981667 | (0.024944) |
| **Naïve Bayes** | 0.986667 | (0.008498) |

## K-Fold Cross-Validation

* Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.
* The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.
* Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.
* It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

**The general procedure is as follows:**

* Shuffle the dataset randomly.
* Split the dataset into k groups
* For each unique group:
  + Take the group as a hold out or test data set
  + Take the remaining groups as a training data set
  + Fit a model on the training set and evaluate it on the test set
  + Retain the evaluation score and discard the model
* Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

# **Precision and Recall**

As there is a high class imbalance in the data, we need to check for the recall and precision to judge the model which is best for our data.

## Recall

* **Recall** attempts to answer the following question:
* What proportion of actual positives was identified correctly?
* Mathematically, recall is defined as follows:



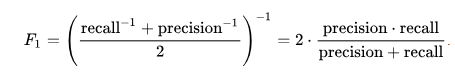
## Precision

* **Precision** attempts to answer the following question:
* What proportion of positive identifications was actually correct?
* Precision is defined as follows:



**F1 Score:**

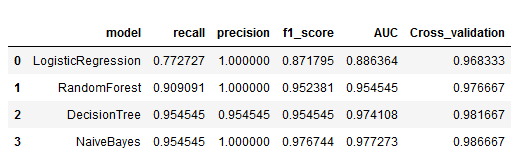
* **F1 score** combines precision and recall relative to a specific positive class.
* The **F1 score** can be interpreted as a weighted average of the precision and recall, where an **F1 score** reaches its best value at 1 and worst at 0.



**Since our aim is to identify customer churn and prevent it. Our objective should be to correctly identify all the customers who are likely to churn i.e to maximise the recall and precision value.**

Below table gives the accuracy metrics for the above mentioned classification algorithms.

We can infer from the accuracy metrics that Naïve bayes Algorithm is giving the best accuracy score along with High Recall, High Precision thus a high f1\_score. The cross validation score is also the maximum for Naïve Bayes Algorithm with a standard deviation of just 0.8%.



**Conclusion**

**We finally chose the model with a relatively higher recall than precision, since the aim is to prevent customer churn.**

**Naïve Bayes Algorithm performed the best.**

**Why Naïve Bayes over other algorithms?**

* High Cross Validation Accuracy score.
* High Recall and Precision value.
* Model did not over-fit or under-fit.
* Assumption of independence of features holds true.
* High Performance of Naïve Bayes model and performs pretty well.
* Our data set was small and Naïve Bayes performs well with categorical features.

**The churn can be predicted by identifying the type of customers and maybe rolling out offers that provide a cheaper tariff with high reliability so that customers are happy and content with the price they are paying for the service and do not go to any other service provider.**

**REFERENCES**

* Customer attrition. Retrieved from http://en.wikipedia.org/wiki/Customer\_attrition, on, February 25, 2011.
* Predictive analytics, Retrieved from http://en.wikipedia.org/wiki/Predictive\_analytics on April 14, 2011.
* Predictive modeling, Retrieved from http://en.wikipedia.org/wiki/Predictive\_modelling, on March 17, 2011.
* R. Mattison, The Telco Churn Management Handbook, 2001.
* Churn Analysis. (n.d.). Retrieved from http://www.ambarasoft.com/researchservices/churnanalysis. Html.
* I. H. Witten, E. Frank. Data Mining Practical Learning Tools and Techniques, Morgan Kaufmann Publishers, 2005.
* Supervised learning Retrieved from http://en.wikipedia.org/wiki/Supervised\_learning, retrieved on April 16, 2011.
* K. B. Oseman, S.B.M. Shukor, N. A. Haris, F. Bakar, “Data Mining in Churn Analysis Model for Telecommunication Industry”, Journal of Statistical Modeling and Analytics, Vol. 1 No. 19-27, 2010.
* T. Mutanen, Customer churn analysis – a case study, Technical Report, Retrieved from, http://www.vtt.fi/inf/julkaisut/muut/2006/customer\_churn\_c ase\_study.pdf, April 12, 2014.
* S.V. Nath, Customer Churn Analysis in the Wireless Industry: A Data Mining Approach, Technical Report, retrieved from http://download.oracle.com/owsf\_2003/40332.pdf, April 14, 2014.
* M. Richeldi and A. Perrucci, “Churn Analysis Case Study”, Technical Report, Telecom Italia Lab, Italy, retrieved from http://sfb876.tudortmund. de/PublicPublicationFiles/richeldi\_perrucci\_2002b.pdf, April 12, 2014.