Sriganesh Balamurugan – 11915001   
Raghu Punnamraju – 11915010

Anmol More – 11915043

1. ***By visualizing the data (without any statistical test), can you claim that Wall’s belief about the dependence of the churn rates on customer age is supported?***

**Assumption –** We have dropped the customers for which we don’t have any data, ie. None of the feature values are present/ just marked as zero in data file. We had 765 such rows.

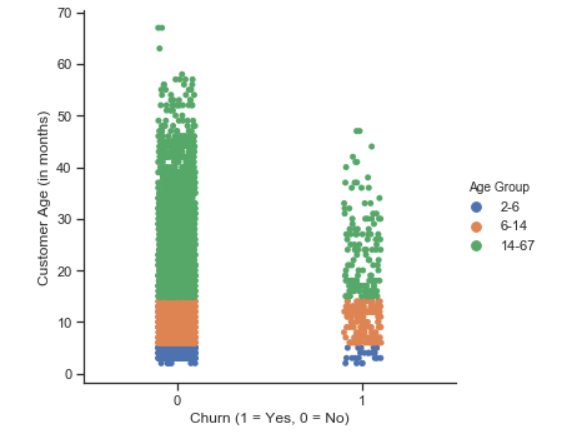
Based on the graphs (Exhibit 1, Exhibit 2) below we can see that, the churn rate is highest in the 6-14 age group i.e. 7.23% And in other Age Groups it is 2.76% and 5.51%.

Yes, based on this data we can say that Wall’s belief hold true. However, to be certain statistical test is required.

**Exhibit 1 : Table of customers churning in each age group**

|  |  |  |  |
| --- | --- | --- | --- |
| **Age Group** | **Churn (1 = Yes, 0 = No)** | **counts** | **percentage** |
| 14-67 | 0 | 2262 | 94.49 |
|  | 1 | 132 | 5.51 |
| 2-6 | 0 | 881 | 97.24 |
|  | 1 | 25 | 2.76 |
| 6-14 | 0 | 2117 | 92.77 |
|  | 1 | 165 | 7.23 |

**Exhibit 2 : Percentage of customers churning in each age group**



* 1. ***2. Run a single regression model that best predicts the probability that a customer leaves. Here, a single regression model means one regression model with all the data (without subsampling). It doesn’t mean a simple regression model with a single variable.***

**Assumptions** –

* Customer Age (in months) = 0 are not useful for this analysis, as we don’t have any data about them (all feature values are 0). So we have kept that aside, and not considered them for regression
* CHI Score 0-1, refers to (CHI Score of month 0 – CHI Score of month 1) . So, negative CHI Score 0-1 would mean customer’s CHI score decreased as compared to month 1 (1 month back). Same assumption follows for other variables with similar naming structure - Support Cases 0-1, SP 0-1, Logins 0-1, Blog Articles 0-1, Views 0-1, Days Since Last Login 0-1

**Single regression model suffers from disbalanced dataset.** Accuracy of Logistic regression classifier : 0.94

* 1. **Exhibit – 3 : Classification matrix of logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted (Not Churn) | Predicted (Churn) |  |
| Actual (Not Churn) | 5260 | 0 |  |
| Actual (Churn) | 322 | 0 |  |

* + 1. ***a. What is the predicted probability that customer 672 will leave between December 2011 and February 2012? Is that high or low? Did that customer actually leave?***
    2. ***Predicted probability of customer 672 will leave - .035. No customer did not actually leave***
  1. ***b. What about customers 354 and 5203?***
  2. Both of them did not leave, probabilities below
  3. ***Exhibit 4***

|  |  |  |
| --- | --- | --- |
| **Customer ID** | **Probability (Not Churn)** | **Probability (Churn)** |
| 354 | 0.95 | 0.05 |
| 5203 | 0.96 | 0.04 |

* 1. ***3. How sensible is the approach with a single regression model? Can you suggest a better approach? Provide updated estimates of probabilities that customers 672, 354 and 5,203 will leave. What factors contribute the most to the predicted probabilities that these customers will leave ?***

**Single regression model predicts all customers as – Do not churn, which is not sensible.**

This **approach suffers from imbalanced dataset,** where even after cleaning the dataset size is :

5260 – Not churn

322 – Churn

**Better Approach** –

* Use oversampling to balance the dataset
* Split the oversampled dataset in train and test for model tuning and test.
* Run random forest regressor on train set and test accuracy on test set. Use cross validation for model tuning.

Probabilities of customers churning –

|  |  |  |
| --- | --- | --- |
| **Customer ID** | **Probability (Not Churn)** | **Probability (Churn)** |
| 672 | 0.76 | 0.24 |
| 354 | 0.8 | 0.2 |
| 5203 | 0.76 | 0.24 |

Based on above probabilities, we can say none of these three customers will churn.

As explained in python code, Factors contributing most to the predicted probabilities in descending order and percentages of contribution towards model –

|  |  |
| --- | --- |
| Days Since Last Log in 0-1 | 0.174212 |
| CHI Score Month 0 | 0.172156 |
| CHI Score 0-1 | 0.127135 |
| Views 0-1 | 0.125404 |
| Logins 0-1 | 0.121551 |
| Blog Articles 0-1 | 0.106423 |
| Support Cases 0-1 | 0.049683 |
| Support Cases Month 0 | 0.037506 |
| SP 0-1 | 0.026668 |
| Group\_6-14 | 0.016764 |
| Group\_2-6 | 0.014517 |
| Group\_14-67 | 0.014142 |
| SP Month 0 | 0.013839 |