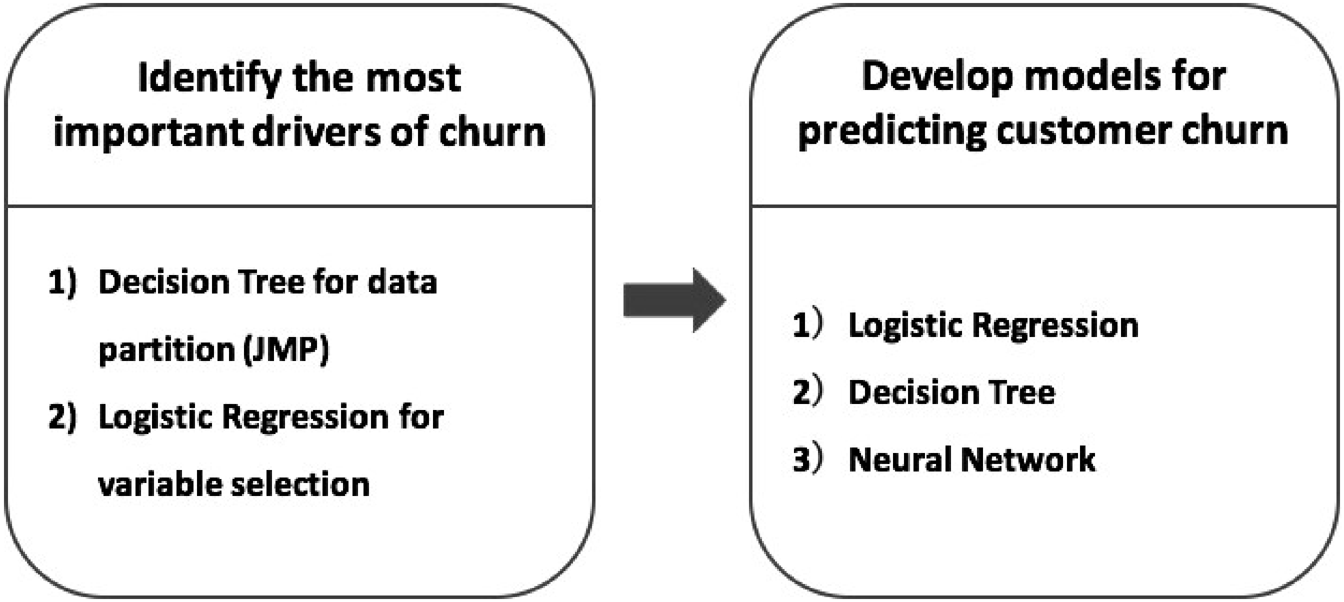
# Overview: Cell2Cell Company

- Over 70,000 rows, with 78 different variables
- Variables we chose to use and what they stand for
    - Eqpdays: *Number of days of the current equipment*
    - Months: *Months in Service*
    - Mou: *Mean monthly minutes of use*
    - Recchrge: *Mean total recurring charge*
    - Retcalls: *Number of calls previously made to retention team*
- Why we chose to use this dataset...

# Business Objectives

1. Create model to understand effect of variables on churn
2. Predict churn through comparison of multiple models
3. Develop plan to reduce churn based on model
4. Utilize model to increase profitability

# Methodology



# Decision Tree - Variable Selection

- Data: Cell2Cell Original Data (71047 rows)
- Dependant variable: Churn
- Independant variables- All (reject calibrate, customer, churndep, and csa)
- Objective:
  - To know the significant variables
- Method: JMP-Analyze-Partition-Decision Tree
- Result:
  - AICC: 89357
- Conclusion:
  - eqpdays,months,mou,recchrge,retcalls are top 5 drivers to Churn

All Rows

| | |
|---|---|
| Count | 71047 |
| Mean | 0.2900756 |
| Std Dev | 0.4538002 |

Candidates

| Term | Candidate SS | | LogWorth | Cut Point |
|---|---|---|---|---|
| eqpdays | 337.6191030 | * | 488.0969074 | 305 |
| months | 248.1093350 | | 355.5767554 | 11 |
| mou | 109.7735953 | | 153.8426187 | 0.5 |
| recchrge | 74.9078866 | | 103.7604833 | 34.99 |
| retcalls | 78.6467017 | | 101.6264279 | 1 |
| retcall | 78.6467017 | | 87.9959719 | 1 |
| changem | 51.9622830 | | 71.0086488 | 0.25 |
| webcap | 56.0345404 | | 62.8968694 | 1 |
| incalls | 45.5566330 | | 61.9036474 | 1 |
| peakvce | 45.1435682 | | 61.3172103 | 9.33 |
| changer | 43.8369006 | | 59.4626987 | 406.93 |
| opeakvce | 42.6951745 | | 57.8430419 | 0.67 |
| creditde | 45.9815919 | | 51.7539859 | 1 |
| custcare | 37.6880047 | | 50.7488741 | 1.33 |
| outcalls | 35.8329253 | | 48.1246765 | 4 |
| unansvce | 35.7179679 | | 47.9621365 | 0.33 |
| mourec | 26.9863838 | | 35.6480393 | 0.07 |
| dropblk | 26.8615980 | | 35.4725790 | 0.33 |
| uniqsubs | 24.7947462 | | 32.6637977 | 2 |
| models | 24.3460871 | | 32.0400040 | 2 |
| retaccpt | 21.6752533 | | 27.4951138 | 1 |
| phones | 20.4669115 | | 26.8090791 | 2 |
| setprc | 19.1975976 | | 24.9505436 | 9.99 |
| revenue | 18.8660577 | | 24.2735812 | 28.46 |

# Logistic Regression - Variable Selection

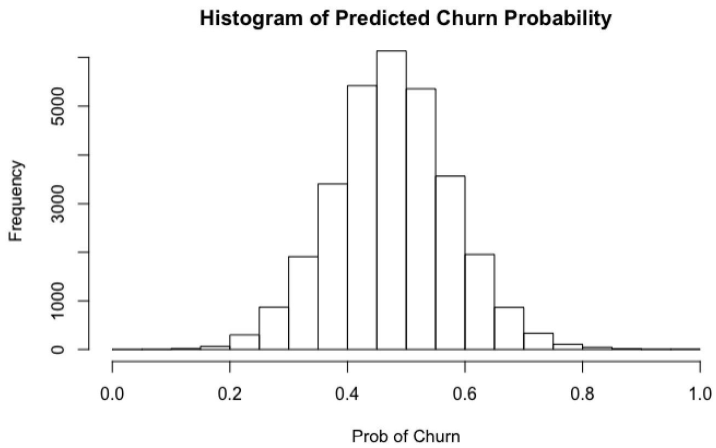
- Data: Entire dataset
- Dependant variable- Churn
- Independant variables- All (except calibrate, customer, churndep, and csa)
- Objective:
  - To know the significant variables
- Method: R function- glm()
- Result:
  - AICC: 52909
- Code: `glm.c2c <- glm(churn ~ .-calibrat-churndep, family=binomial(link='logit'), data=c2c)`
- Conclusion:
  - AICC for decision tree higher

```
Coefficients: (2 not defined because of singularities)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.722e+00  8.820e-01   4.220 2.44e-05
revenue         1.720e-03  8.101e-04   2.123 0.033763
mou            -2.703e-04  5.094e-05  -5.306 1.12e-07
recchrge       -2.817e-03  9.048e-04  -3.113 0.001852
directas       -3.410e-03  6.096e-03  -0.559 0.575892
overage         8.332e-04  2.844e-04   2.929 0.003398
roam            7.180e-03  2.110e-03   3.402 0.000668
changem        -5.104e-04  5.451e-05  -9.365  < 2e-16
changer         2.343e-03  3.743e-04   6.260 3.85e-10
dropvce         1.130e-02  7.307e-03   1.546 0.122043
blckvce         6.618e-03  7.213e-03   0.917 0.358939
unansvce        1.083e-03  4.694e-04   2.307 0.021053
custcare       -5.940e-03  2.615e-03  -2.271 0.023135
threeway       -3.205e-02  1.163e-02  -2.755 0.00587
```

# Logistic Regression - Prediction

- Data: Training data (40,000 rows), Testing data (31047 rows)
- Dependant variable- Churn
- Independant variables- All (except calibrate, customer, churndep)
- Objective:
  - To predict churn
- Result:
  - **Precision:** Fraction of relevant instances among the retrieved instances= 2.81%
  - **Recall**: Fraction of the total amount of relevant instances that were actually retrieved correctly=58.81%

| | Actual Churn | | |
|---|---|---|---|
| Predicted Churn | 0 | 1 | Row Total |
| 0 | 17881<br>0.987 | 241<br>0.013 | 18122<br>0.597 |
| 1 | 11901<br>0.972 | 345<br>0.028 | 12246<br>0.403 |
| Column Total | 29782 | 586 | 30368 |



# Logistic Regression - Prediction

- We chose to use the **5 variables** that we concluded were the best in our Decision Tree Analysis
- Same objectives and data as before but different variables

1.) Eqpdays: *Number of days of the current equipment*

2.) Months: *Months in Service*

3.) Mou: *Mean monthly minutes of use*

4.) Recchrge: *Mean total recurring charge*

5.) Retcalls: *Number of calls previously made to retention team*

***Precision = 2.62%***

***Recall = 59.96%***

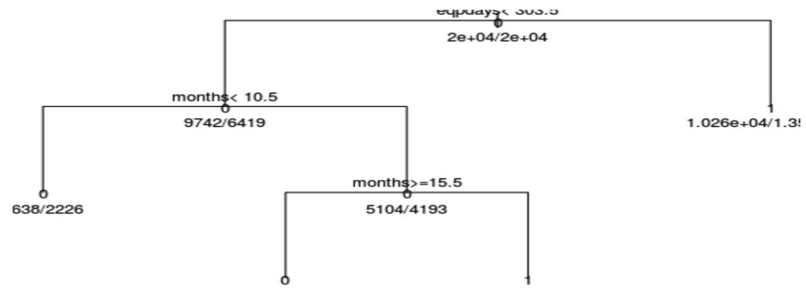
| | Real Response | | |
|---|---|---|---|
| Predicted Response | 0 | 1 | Row Total |
| 0 | 18209<br>0.985 | 277<br>0.015 | 18486<br>0.597 |
| 1 | 12158<br>0.974 | 327<br>0.026 | 12485<br>0.403 |
| Column Total | 30367 | 604 | 30971 |

# Decision Tree - Prediction

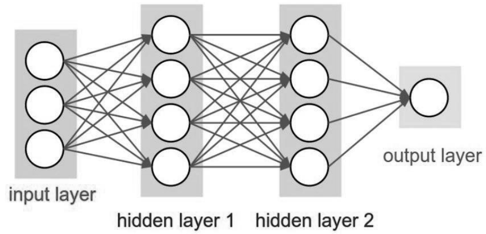
- Data: Training data (40,000 rows), Testing data(31047 rows)
- Dependant variable- Churn
- Independent variables- eqpdays, months, mou, recchrge, retcalls
- Objective:
  - To predict churn
- Method: R package-rpart
- Result:
  - ***Precision=2.5%***
  - ***Recall=42.8%***

| | Actual Churn = 0 (Not Churn) | Actual Churn = 1 (Churn) | Total |
|---|---|---|---|
| Prediction Churn = 0 (Model predicts not churn) | 12,836 | 156 | 12,992 |
| Prediction Churn = 1 (Model predicts churn) | 17,602 | 453 | 18,055 |
| Total | 30,438 | 609 | 31,047 |

**Classification Tree for Cell2Cell Data**



# Neural Network - Prediction




**R package -** ***neuralnet***

- A machine learning that mimics the functioning of ***human brain*** and consists of a number of neurons that continuously interact with each other

**Data Manipulation**

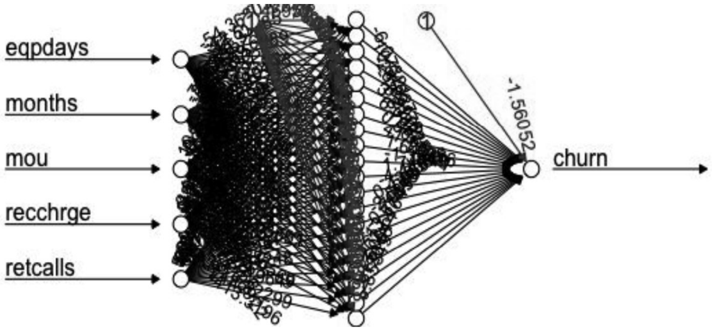
- Dependent variables: ***churn***
- Independent variables: 5 variables with the highest logworth from decision tree model
  - ***eqpdays, months, mou, recchrge, retcalls***
- Remove rows with missing values
- Normalize the data to prevent a particular variable affecting the prediction due to its large numeric value range
  - ***Training data : 39,859 rows, churn rate is approx. 50%***
  - ***Testing data: 30,971 rows, churn rate is approx. 2%***

**Model Processing**

- Randomly select 5,000 observations from the testing data to build the neural network model
- Train the neural network by testing different number of hidden layers: 20 or 10*2
- Predict churn behavior with the neural network
- Create a cross table to judge the quality of the predictions

# Neural Network - Prediction

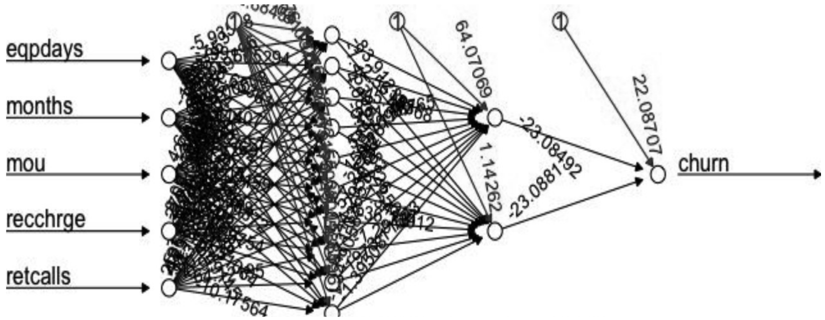
## Model 1– Hidden layers: 20




| Predict | Actual 0 | 1 | Row Total |
|---|---|---|---|
| 0 | 16753 | 249 | 17002 |
| 1 | 13614 | 355 | 13969 |
| Column Total | 30367 | 604 | 30971 |

Precision: 355/13969 = 2.54%
Recall: 355/604 = 58.8%

## Model 2– Hidden layers: 10 * 2




| Predict | Actual 0 | 1 | Row Total |
|---|---|---|---|
| 0 | 15719 | 214 | 15933 |
| 1 | 14648 | 390 | 15038 |
| Column Total | 30367 | 604 | 30971 |

Precision: 390/15038 = 2.59%
Recall: 390/604 = 64.6%

# Comparison of the 4 Models

| | Decision Tree (5 Variables) | Logistic Regression (All Variables) | Logistic Regression (5 Variables) | Neural Network (5 Variables) |
|---|---|---|---|---|
| **Precision Rate** | 2.51% | 2.81% | 2.62% | 2.59% |
| **Recall Rate** | 42.8% | 58.81% | 59.96% | 64.6% |

- The decision tree with several branches predict poorly.
- Neural network model with 5 variables predicts well with highest recall rate.
- Compared with logistic regression model with all variables, logistic model with 5 selected variables has slightly lower precision rate but higher recall rate.
- Overall, logistic regression model and neural network are both acceptable models to predict customer churn behavior.

# Recommendations

1. ***Equipment Ownership Duration*** -
   a. Create segments and market best fitting phones before critical replacement threshold
   b. Ensure Cell2Cell has most popular selection of devices with robust fulfillment channel.

2. ***Service Months*** -
   a. Increase outreach to flight-risk clients as months of service reach critical point.
   b. Ensure these clients are receiving segment specific marketing.

3. ***Avg. Minutes of Use* -** Those that use their phones most look for a best rate. Ensure competitive plans for longtime customers that average high monthly usage.

4. ***Recurring Charges* -** Do not wait until clients call to cancel their account to offer better price. Proactively reach out to customers with to offer lower priced plans when available.

5. ***Customer Service Calls* -** Have manager reach out to those with above average calls to company to ensure everything is good and ensure needs are met or exceeded.

# Limitations and Summary

**Limitations:**

1. Low processing power: reduced training dataset for neural network
2. 2% of testing data had relevant variable (churn=1)
3. Highest predicted probability of churn with decision tree is 0.56

**Summary:**

- Decision tree useful for knowing significant variables
- Prediction/Scoring Logistic Regression and Neutmodel best fit
- Recall and Precision can not be increased simultaneously
- Neural network has highest recall rate
- *Number of days of the current equipment, Months in Service, Mean monthly minutes of use, Mean total recurring charge, Number of calls previously made to retention team* are top 5 drivers to Churn