# Predicting At-Risk Customers with Data Science

**James Richards**
General Assembly sfdat26
James.richards@trinet.com

# About TriNet

- **Founded in 1980s as an HR services firm**
    - Outsourced HR: payroll, benefits, workers comp, and more
- **~13,500 clients w/ ~325,000 employees**
    - $600MM in net revenue/year
    - $31BN in payroll and $3BN in health insurance premiums
    - “Most average” client: 20-person tech startup in the Bay Area or New York.
    - Diverse client base: retail, hotels, restaurants, hedge funds, etc…

# Problem: We start every year ~$100MM in the hole

- **~15% “voluntary” churn every year**
  - Controlling for clients going out of business, getting acquired, etc
  - Costs us over $100MM per year
  - We have no idea why.
- **Digging out of the hole <u>sucks</u>**
  - Wall Street expects 20% growth
  - Target: $600MM * 1.20 = $720MM
  - Start: $600MM – 15% = $510MM
  - <u>Need to add $210MM in new revenue just to meet expectations!</u>

# Honestly, we don’t know why clients leave

- **We classify clients manually and reactively**
  - 1 Account Manager : 50-100 clients
  - 1 Director : 6-8 Account Managers
  - Every 2 weeks, Directors classify clients into Green, Yellow, or Red based on subjective impressions of Account Managers
- **No playbook to actively intervene and prevent churn**
  - Only response with Red accounts is to lower price
  - Often, many clients go from Green to Red because they’ve told us they have already decided to leave

# We paid BCG a lot of money to figure it out...

- **According to BCG, churn is correlated with:**
  - Payroll accuracy – % of inaccurate payrolls we run for a client
  - Client growth – % growth in headcount a client experiences in a year
- **Recommendation:**
  - Fix payroll accuracy
  - Focus Account Manager time on clients who are growing quickly

THE BOSTON CONSULTING GROUP

# …but does the BCG "model" make sense?

- **BCG recommendation is essentially a high-bias, low variance model:**
  - Risk of churn increases linearly with growth in headcount and payroll errors
- **Intuitively, is that the best model?**
  - Hedge funds rarely experience any growth in headcount but still churn
  - Highly variant population and variant reasons for churn
  - Often a combination of features (based on exit interviews
- **Is it useful in real life?**
  - Can't fix payroll errors after the fact
  - Can't ungrow a client

## Can we use data science to classify clients as “at risk” before it’s too late?


I'M GOING TO HAVE TO SCIENCE
THE SHIT OUT OF THIS
ARES III

# Our data set: 9 months of client data

- Data set: 9 months of customer data

```
In [7]: clients.Termed.value_counts()
Out[7]: False    7510
        True      722
        Name: Termed, dtype: int64
```

Out[2]:

| | Unnamed: 0 | Company | Company.Name | City | State | Related.Company | Company.Live.Date | NAICS.2 | NAICS.3 | NAICS.4 | ... | Total.Voluntary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 4AW | HRchitect Canada | Frisco | TX | NaN | 1/12/2004 | NaN | NaN | NaN | ... | NaN |
| **1** | 2 | 4B9 | RouteOne - Canada | Farmington Hills | MI | NaN | 2/16/2004 | NaN | NaN | NaN | ... | NaN |
| **2** | 3 | 4BJ | Ekaria LLP - Canada | Redmond | WA | 495 | 1/1/2004 | NaN | NaN | NaN | ... | NaN |
| **3** | 4 | 4BY | MicroVention, Inc. - Canada | Aliso Viejo | CA | NaN | 2/1/2004 | NaN | NaN | NaN | ... | NaN |
| **4** | 5 | 51Z | Signature Control Systems CAN | Sheridan | CO | NaN | 8/1/2006 | NaN | NaN | NaN | ... | NaN |

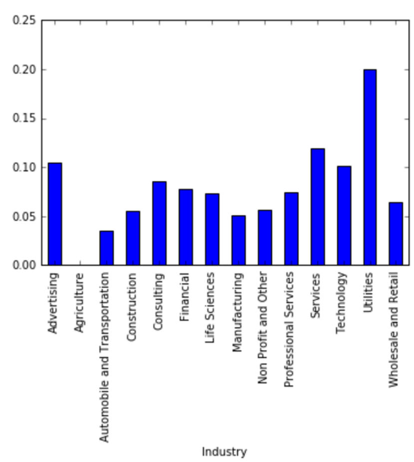
5 rows × 119 columns

# Initial observations/EDA

- Dropped features with null values, added payroll accuracy (1 – (payrolls with error / payrolls))
- No obviously good features
- Low sample size on Industry and State
- Payroll accuracy doesn't at first appear to be the right

```
In [51]: clients.groupby('Termed').Payroll_Accuracy.mean()

Out[51]: Termed
         0    0.990344
         1    0.990116
         Name: Payroll_Accuracy, dtype: float64
```

```
In [52]: clients.groupby('Industry').Termed.mean().plot(kind='bar')

Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x11b6f3210>
```




```
In [53]: clients.Industry.value_counts()

Out[53]: Technology                      2518
         Financial                       1097
         Professional Services           1066
         Advertising                      892
         Services                         572
         Consulting                       548
         Non Profit and Other             423
         Life Sciences                    408
         Wholesale and Retail             325
         Construction                     145
         Manufacturing                    139
         Automobile and Transportation     84
         Utilities                         10
         Agriculture                        4
```

# First cut: Correlation matrix

- Still no standout feature(s)
- Strongest features appear to be:
  - Price Change Amount
  - Involuntary and Voluntary Terminations
  - # of employees
  - # of payrolls (correlated w/ employees)
  - Interactions, esp Benefits and Tax
  - Things to do w/ Payroll have weak correlation (except accuracy, which is negatively correlated w/ termination?)

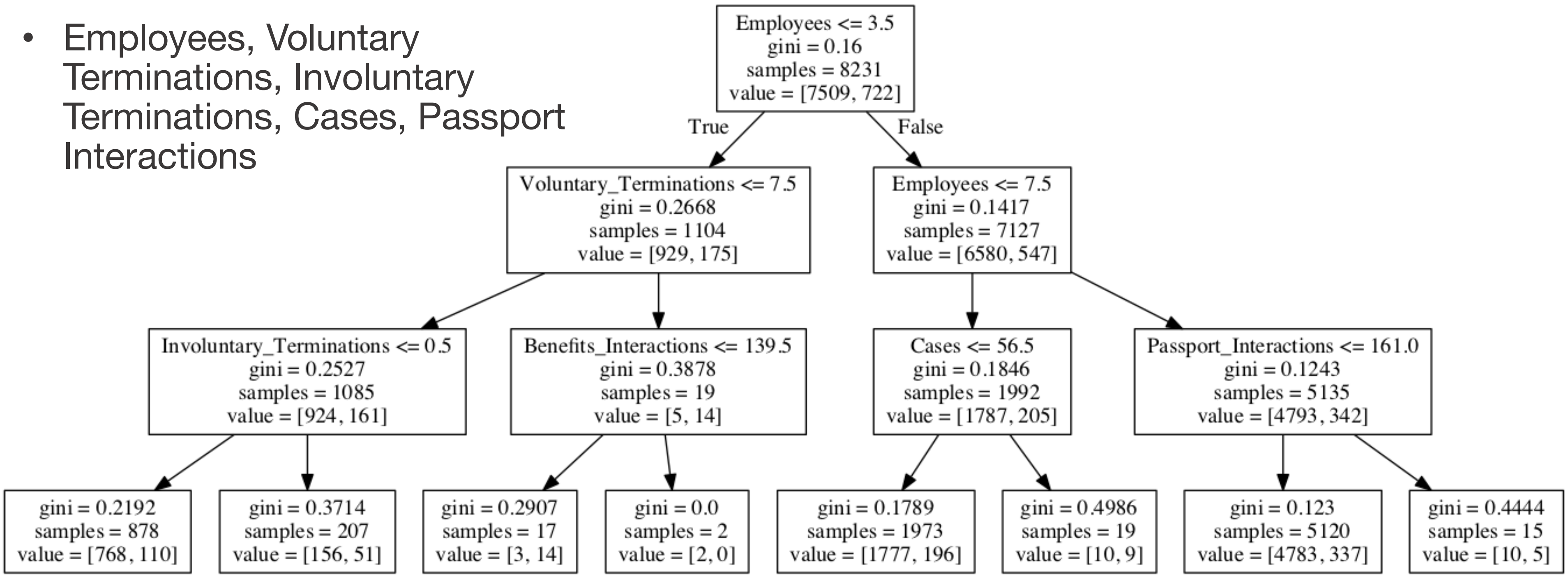
Out[56]:

| | ID | Termed |
|---|---|---|
| ID | 1.000000 | -0.008917 |
| Termed | -0.008917 | 1.000000 |
| Employees | 0.012272 | -0.029586 |
| Expense | -0.022639 | 0.003438 |
| Hire | 0.010906 | -0.004647 |
| Perform | 0.013375 | -0.000076 |
| Price_Change | -0.004628 | -0.012578 |
| Price_Change_Amount | -0.015709 | 0.016474 |
| Cases | -0.000799 | -0.020782 |
| Benefits_Interactions | -0.003811 | -0.025912 |
| Cust_Service_Interactions | -0.000827 | -0.005492 |
| Global_Services_Interactions | 0.012367 | -0.011895 |
| Passport_Interactions | 0.023389 | 0.006831 |
| HR_Interactions | 0.027235 | -0.021010 |
| IT_Enhancement_Interactions | -0.001550 | -0.015972 |
| IT_Maintenance_Interactions | 0.003233 | -0.012364 |
| IT_Problem_Interactions | 0.003578 | -0.021864 |
| Standard_IT_Interactions | -0.006914 | -0.000060 |
| LOA_Interactions | 0.043486 | -0.021172 |
| Legal_Interactions | 0.015489 | -0.006249 |
| Payroll_Interactions | 0.006014 | -0.013408 |
| Product_Interactions | 0.021623 | -0.006197 |
| Risk_Interactions | -0.018837 | -0.002473 |
| Underwriting_Interactions | -0.005785 | -0.010521 |
| Strategic_Service_Interactions | -0.000541 | -0.018742 |
| Tax_Interactions | 0.012212 | -0.019914 |
| Unemployment_Claim_Interactions | -0.003365 | -0.001088 |
| WC_Interactions | 0.024391 | -0.002935 |
| Payrolls | -0.063796 | -0.024160 |
| Payrolls_With_Error | 0.007935 | -0.004213 |
| Voluntary_Terminations | 0.020890 | 0.000463 |
| Involuntary_Terminations | 0.008109 | 0.011889 |
| Total_Terminations | 0.016564 | 0.006370 |
| Payroll_Accuracy | -0.003134 | -0.002023 |

# First model: Decision tree

- 3 layers deep

- Employees, Voluntary Terminations, Involuntary Terminations, Cases, Passport Interactions



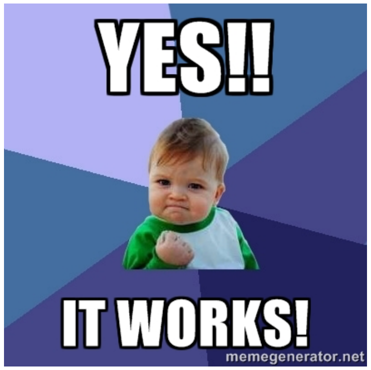

- Terminal nodes not all that pure…

# Second model: KNN

- Fed features into KNN, ran train/test split and cross-val to find optimal K value
- Got to almost 92% accuracy with 7 neighbors

```
In [228]: knn = KNeighborsClassifier(n_neighbors=7)
          knn.fit(features_train, response_train)
          knn.score(features_test, response_test)

Out[228]: 0.918367346938775
```

```
In [189]: feature_cols = ['Employees', 'Price_Change_Amount', 'Involuntary_Terminations', 'Price_Change',
          X = clients[feature_cols]
          y = clients.Termed

          print X.shape
          print y.shape

          (8231, 28)
          (8231,)

In [190]: from sklearn.neighbors import KNeighborsClassifier  # import class

In [191]: knn = KNeighborsClassifier(n_neighbors=3)           # instantiate the estimator

In [192]: knn.fit(X, y)                                       # fit with data

Out[192]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=1, n_neighbors=3, p=2,
                     weights='uniform')

In [193]: knn.score(X, y)

Out[193]: 0.92273113837929777
```



# Summary & next steps

- It is possible to predict accounts who may churn
- Figuring out why clients churn is really hard
- Long, complex relationships over several years
- Payroll accuracy has limited predictive power
- **Next steps:**
  - Go deeper into pricing and change in employees over time
  - Look at correlations between manual classification of risk factors and term
  - Interview Account Managers to see what factors they think are important
  - Try to get more granular feature data (e.g., good interactions vs bad interactions)