A Model to Predict When Employees will take Sick-Leave

Bryan Feeney

Mudano Interview Process

November 17, 2017

Introduction

Task: Predict when team-members are likely to take sick-leave

Introduction

Task: Predict when team-members are likely to take sick-leave

- Sick-Leave Patterns and Indicators
- Proposed Model
- Evaluation
- Engineering

Sick-Leave Patterns & Indicators

Sickness absence from work in the UK[2]

An quantitative analysis of 1.7m sick-leave absences between March and May 2004 recorded as part of the UK Labour-Force Survey

Sick-Leave Patterns & Indicators

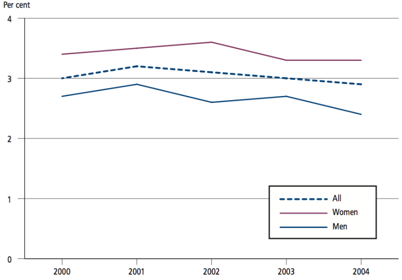
Sickness absence from work in the UK[2]

Incidence : 2.9% of workers are ill ≥ 1 days (average 9/228)

Sick-Leave Patterns & Indicators

Sickness absence from work in the UK[2]

Incidence : 2.9% of workers are ill ≥ 1 days (average 9/228)
Gender : Women are 50% more likely to take a sick day



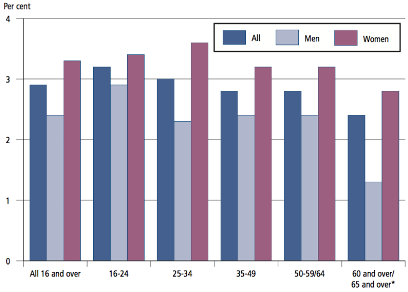
Source: Labour Force Survey

a Proportions of employees who were absent from work for at least one day in the reference week.

Sick-Leave Patterns & Indicators

Sickness absence from work in the UK[2]

Incidence	:	2.9% of workers are ill ≥ 1 days (average 9/228)
Gender	:	Women are 50% more likely to take a sick day
Age	:	Junior staff (16-34) are most likely to be absent



Source: Labour Force Survey

a Proportions of employees who were absent from work for at least one day in the reference week.
** This estimate is based on small sample sizes and is subject to large sampling variability.*

Sick-Leave Patterns & Indicators

Sickness absence from work in the UK[2]

Incidence : 2.9% of workers are ill ≥ 1 days (average 9/228)

Gender : Women are 50% more likely to take a sick day

Age : Junior staff (16-34) are most likely to be absent

Weekday : Midweek absences are most likely

Day:	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Absences:	17,670	19,051	19,115	18,914	17,883	4,610	2,520
Percent:	17.7	19.3	19.0	19.1	18.5	4.1	2.3

Sick-Leave Patterns & Indicators

Sickness absence from work in the UK[2]

Incidence : 2.9% of workers are ill $\geq$ 1 days (average 9/228)

Gender : Women are 50% more likely to take a sick day

Age : Junior staff (16-34) are most likely to be absent

Weekday : Midweek absences are most likely

Profession : Industry, role and Public/Private are predictive

Location : Londoners are more likely to be sick (3.1%) compared to non-Londoners (2.2%)

Sick-Leave Patterns & Indicators

Risk of future sickness absence in frequent and long-term absentees[7]

- **Incidence** : 2.9% of workers are ill ≥ 1 days (average 9/228)
- **Gender** : Women are 50% more likely to take a sick day
- **Age** : Junior staff (16-34) are most likely to be absent
- **Weekday** : Midweek absences are most likely
- **Profession** : Industry, role and Public/Private are predictive
- **Location** : Londoners are more likely to be sick (3.1%) compared to non-Londoners (2.2%)
- **Recurrence** : Frequent absentees have a higher risk of future absences, and long-term absences

Sick-Leave Patterns & Indicators

Trends and seasonality in absenteeism[1]

Incidence	: 2.9% of workers are ill ≥ 1 days (average 9/228)
Gender	: Women are 50% more likely to take a sick day
Age	: Junior staff (16-34) are most likely to be absent
Weekday	: Midweek absences are most likely
Profession	: Industry, role and Public/Private are predictive
Location	: Londoners are more likely to be sick (3.1%) compared to non-Londoners (2.2%)
Recurrence	: Frequent absentees have a higher risk of future absences, and long-term absences
Season	: Absences peak in Winter months, and are minimal in Summer months

Sick-Leave Patterns & Indicators

Final Feature Set

- Gender (via name) : Categorical (3)
- Project role (proxy for age) : Categorical (?)
- Weekday : Categorical (7)
- Individuals' sick-leave days in the last 12 months : Numeric
- Teams' sick-leave days in the last four weeks : Numeric
- Time of Year: encoded using a number of Fourier terms ? $\times$ numeric
  - For day d produce a 2-vector $\left(cos(2\pi\frac{d}{365}), sin(2\pi\frac{d}{365})\right)$.

Ignore

- Location

Baseline Model

Baseline Model

- Average per-month absenteeism by gender

Baseline Model

Baseline Model

- Average per-month absenteeism by gender

Prediction:

- For a given month m
- When individual u's absentee rate is a_{um}
- In which a 15-day sprint occurs

Baseline Model

Baseline Model

- Average per-month absenteeism by gender

Prediction:

- For a given month m
- When individual u's absentee rate is a_{um}
- In which a 15-day sprint occurs
- $\Pr(\text{absences} = y | \text{sprint} = s, \text{month} = m, \text{user} = u) = \text{Bin}\left(a_{um}, 15\right)$

Baseline Model

Baseline Model

- Average per-month absenteeism by gender

Prediction:

- For a given month m
- When individual u's absentee rate is a_{um}
- In which a 15-day sprint occurs
- Expected number of absences for an individual u is $15a_{um}$

Baseline Model

Baseline Model

- Average per-month absenteeism by gender

Prediction:

- For a given month m
- When individual u's absentee rate is a_{um}
- In which a 15-day sprint occurs
- The expected number of absences in a team is $T.15\left(\frac{1}{T}\sum_{u\in T} a_{um}\right)$
- The variance is $T\bar{a}_m\left(1-\bar{a}_m\right)-\sum_{u\in T}\left(a_{um}-\bar{a}_m\right)$ (see [5])

Custom Model

Logistic Regression; L2-Regularization via Spherical Gaussian prior
Denote the variable extracted from a record as $\phi(x)$

$$
\begin{aligned}
p(y|x) &= \sigma(\mathbf{w}^\top \phi(x)) \\
\mathbf{w} &\sim \mathcal{N}\left(\mathbf{0}, \tau^{-1} I\right) \\
\tau &\sim \mathcal{G}\left(a_0, b_0\right)
\end{aligned}
$$

Use Probit trick[3] for prediction

$$
\begin{aligned}
p(y^*|x^* y, X) &= \int \sigma(\mathbf{w}^\top \phi(x^*)) p(\mathbf{w}|y, X) d\mathbf{w} \\
&\approx \sigma(\kappa(\phi(x^*)\Sigma_N^{(w)} \phi(x^*)) \quad \mathbf{w}_{\mathrm{MAP}}^\top \phi(x^*)) \\
\kappa(\alpha^2) &= (1 + \pi\alpha^2/8)^{1/2}
\end{aligned}
$$

Custom Model : Expediting Training 1

- Inference in Bayesian logistic classically involves a second-order Taylor expansion around the posterior mode to handle the model's non-conjugacy: the Laplace approximation

Custom Model : Expediting Training 1

- Inference in Bayesian logistic classically involves a second-order Taylor expansion around the posterior mode to handle the model's non-conjugacy: the Laplace approximation
- This involves evaluating and inverting the Hessian at every iteration

Custom Model : Expediting Training 1

- Inference in Bayesian logistic classically involves a second-order Taylor expansion around the posterior mode to handle the model's non-conjugacy: the Laplace approximation
- This involves evaluating and inverting the Hessian at every iteration
- The Bohning bound[4] replaces the true Hessian with an upper bound: $\hat{\mathrm{H}} = \frac{1}{2}\left(I - \frac{1}{2}\mathbf{1}\mathbf{1}^\top\right), \quad \hat{\mathrm{H}}^{-1} = 2\left(I + \mathbf{1}\mathbf{1}^\top\right)$

Custom Model : Expediting Training 1

- Inference in Bayesian logistic classically involves a second-order Taylor expansion around the posterior mode to handle the model's non-conjugacy: the Laplace approximation
- This involves evaluating and inverting the Hessian at every iteration
- The Bohning bound[4] replaces the true Hessian with an upper bound: $\hat{\mathrm{H}} = \frac{1}{2}\left(I - \frac{1}{2}\mathbf{1}\mathbf{1}^\top\right), \quad \hat{\mathrm{H}}^{-1} = 2\left(I + \mathbf{1}\mathbf{1}^\top\right)$
- This reduces the time per iteration, and guarantees convergence, but increases the number of iterations involved

Custom Model : Expediting Training 2

The dataset is heavily skewed, so focus only on positive examples and skip negative ones, using weighted stochastic gradient descent[6][supporting material]

SGD

$$\mathbb{E}_{p^*}[D \cdot f(x)] \approx \frac{D}{S}\sum_s f(x_s), \qquad x_s \sim p^* \qquad p^*(x) = \frac{1}{D}\sum_d 1_{x=x_d}$$

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} \frac{D}{S}\sum_s f(x_s)$$

Custom Model : Expediting Training 2

The dataset is heavily skewed, so focus only on positive examples and skip negative ones, using weighted stochastic gradient descent[6][supporting material]

Weighted SGD
Use a proposal distribution $q(x)$ selecting negative examples with small probability ϵ and positive examples probability $1-\epsilon$

$$
\begin{aligned}
\mathbb{E}_{p^*}[D \cdot f(x)] &= \mathbb{E}_{p^*}[D \cdot f(x)\frac{q(x)}{q(x)}] \\
&\approx \frac{D}{S}\sum_s f(x_s)\frac{p^*(x_s)}{q(x_s)}q(x_s) \\
&\approx \mathbb{E}_{q^*}[D \cdot f(x)\frac{p^*(x)}{q(x)}]
\end{aligned}
$$

Custom Model : Expediting Training 2

The dataset is heavily skewed, so focus only on positive examples and skip negative ones, using weighted stochastic gradient descent[6][supporting material]

Weighted SGD
Use a proposal distribution $q(x)$ selecting negative examples with small probability ϵ and positive examples probability $1 - \epsilon$

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} \left(\frac{D^{-}}{S\epsilon} \sum_{s} f(x_s^{-}) + \frac{D^{+}}{S(1-\epsilon)} \sum_{s} f(x_s^{+}) \right)$$

Custom Model : Open Questions

1. Do Sharktower customers' behaviours match the Labour-Force Survey
2. Are there any interactions between features
3. How many Fourier terms are needed to model seasonality
4. Does the Bohning bound interact poorly with SGD

Evaluation

- Incidence is too low for simple accuracy, so use ranking metrics
- Ranking metrics include
  - Precision at M: what proportion of the top M ranked employees were in fact absent
  - Recall at M: what proportion of all absent employees (clamped at M) were in the top M
- Can also compare actual team absences per month with expected team absences

Engineering

Resources

- **Tools** : Python stack: Numpy, Scipy, Scikit-Learn
- **Data (Internal)** : Full list of absences for all members of all teams of all clients with forenames, team IDs and project roles
- **Data (External)** : A name to gender database: likely to start with: http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/

Sprint Work:

- **Week 1** : Acquiring and cleaning data, name-to-gender model, evaluation of baseline model
- **Week 2** : Initial PoC of custom model on acquired dataset, comparisons with baseline
- **Week 3** : Refinements of baseline model, discussion with engineering for productionization if suitable.

Questions

Questions

References I

[1] Ernest B. Akyeampong.
Trends and seasonality in absenteeism.
Perspectives on Labour and Income, 8(6):13–15, 2007.

[2] Catherine Barham and Nasima Begum.
Sickness absence from work in the UK.
Labour Market Trends, (April):149–157, 2005.

[3] Christopher M Bishop.
Pattern recognition and machine learning.
Springer-Verlag, New York, 2006.

[4] Dankmar Bohning.
Multinomial Logistic Regression Algorithm.pdf.
Annals of the Institute of Statistical Mathematics, 40(4):644–6648, 1988.

[5] Zvi Drezner and Nicholas Farnum.
A generalized binomial distribution.
Communications in Statistics - Theory and Methods, 22(11):3051–3063, 1993.

[6] Prem K Gopalan and David M Blei.
Efficient discovery of overlapping communities in massive networks.
Proceedings of the National Academy of Sciences of the United States of America, 110(36):14534–9, 2013.

[7] Petra C. Koopmans, Corné A.M. Roelen, and Johan W. Groothoff.
Risk of future sickness absence in frequent and long-term absentees.
Occupational Medicine, 58:268–274, 2008.