Machine Learning in the real world

I am ***Javier***

Machine learning for Insurance

Machine Learning for Retail (jaggu.com)
B2B ML services (growthintel.com)

DAWN Project: **D**eep Learning to **A**nalyze **W**ebb-detected **N**ascent-galaxies

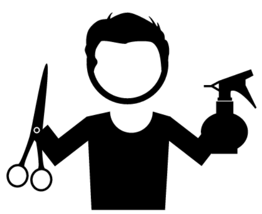
Simply Business

Small businesses want insurance

Go to our website or call center

Get quote from insurers in our panel

1 Simply Business Rating Engine



Please tell us about your business

What is your specific business / trade?

e.g. Builder, Painter and decorator, or Plumber

Or choose from the full list of occupations

Do you have a secondary business activity / secondary trade? ?

Yes | No

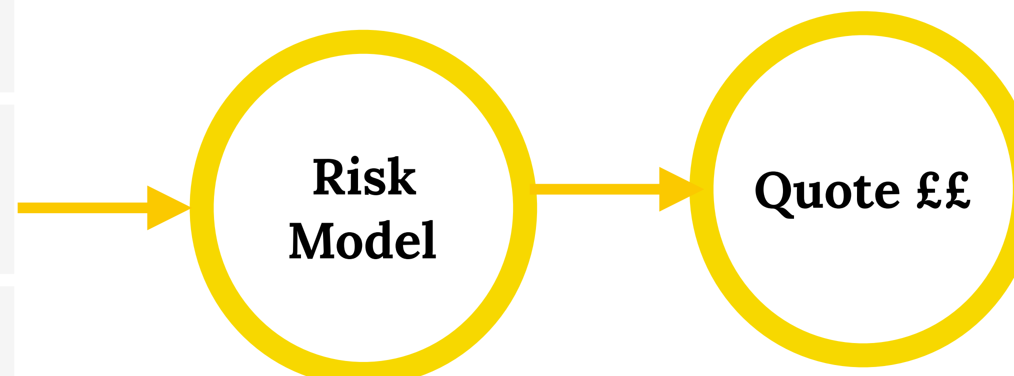
What is your business postcode?

How many years have you been running your own business in this industry? ?

Please select

Which of these categories best describes your business?

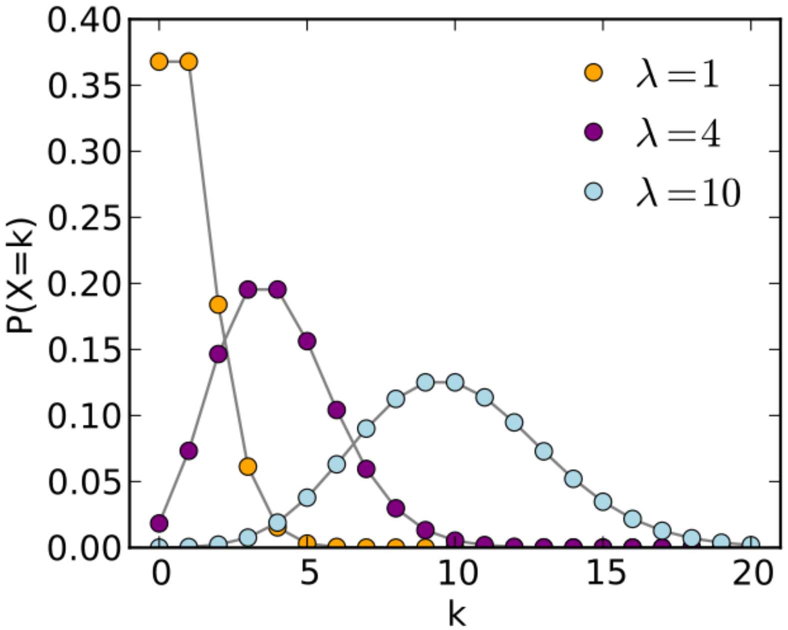
Sole trader | Partnership | Ltd company

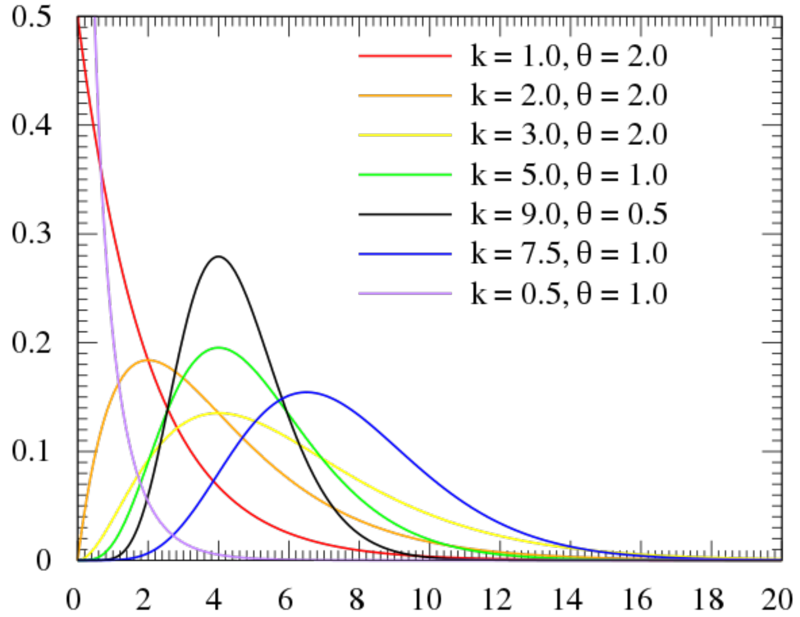


! Traditional Risk Models

Premium = *f*(frequency, severity)

- **Frequency** – modeled with a **Poisson** distribution (or negative binomial distribution if data is over-dispersed
- **Severity** – can be modeled with a **Gamma** distribution
- Data is model with a **GLM** (although some moving on to GBMs)





Traditional Risk Models

- The conditional variance is equal to the conditional mean (solved with a Negative Binomial Distribution)
- The occurrence of one event does not affect the probability that a second event will occur – events occur independently
- The rate at which events occur is constant

Simply Business ML Risk Model

Towards an automated process

2 Simply Business ML Risk Model

0

Data Collection

Redshift+SQL

1

Data Processing

Customers are arrays of 50-150 numbers

2

Classifier + Optimizer + Explainer

lightGBM + hyperopt + LIME

3

Calibration

Prophet + adjusting function

2 SB ML Risk Model: *lightGBM*

- **GOSS**: Gradient-based One-Side Sampling.

  keeps all the instances with large gradients and performs random sampling on the instances with small gradients.

- **EFB**: Exclusive Feature Bundling

  In a sparse feature space, many features are mutually exclusive. One can bundle exclusive features into a single feature (NP-Hard).

- **Natural Treatment of Categorical Features**

  Split on a categorical feature by partitioning its categories into 2 subsets. To find the optimal partition *Light*GBM sorts the histogram according to its accumulated values (*sum_gradient / sum_hessian*) and then finds the best split on the sorted histogram.

2 SB ML Risk Model: *hyperopt*

- **TPE**: tree-structured Parzen estimator (TPE): TPE **models $p(y)$ and $p(x|y)$**

  Models $p(x|y)$ by replacing the distributions of the configuration prior with non-parametric densities.

- The configuration space is described using uniform, log-uniform, quantized log-uniform, and categorical variables.

- The TPE algorithm makes the following replacements: uniform → truncated Gaussian mixture, log-uniform → exponentiated truncated Gaussian mixture, categorical → re-weighted categorical

- Using different observations $\{x(1), \ldots, x(k)\}$ in the non-parametric densities, these substitutions represent a learning algorithm that can produce a variety of densities over the configuration space $\mathcal{X}$

Some Code...

Objective Function

```
def lgb_objective(params):

    lgb_objective.i+=1

    # hyperopt casts as float
    params['num_boost_round'] = int(params['num_boost_round'])
    params['num_leaves'] = int(params['num_leaves'])

    # need to be passed as parameter
    params['is_unbalance'] = True
    params['verbose'] = -1
    params['seed'] = 1

    lgtrain = lgb.Dataset(X_train, label=y_train)
    cv_result = lgb.cv(
        params,
        lgtrain,
        num_boost_round=params['num_boost_round'],
        metrics='binary_logloss',
        nfold=3,
        stratified=True,
        early_stopping_rounds=20)

    error = cv_result['binary_logloss-mean'][-1]
    print("INFO: iteration {} error {:.3f}".format(lgb_objective.i, error))

    return error
```

lightGBM native methods

parameter space

```
lgb_parameter_space = {
    'learning_rate': hp.uniform('learning_rate', 0.01, 0.5),
    'num_boost_round': hp.quniform('num_boost_round', 50, 500, 50),
    'num_leaves': hp.quniform('num_leaves', 30,1024,5),
    'min_child_weight': hp.quniform('min_child_weight', 1, 50, 2),
    'colsample_bytree': hp.uniform('colsample_bytree', 0.5, 1.),
    'subsample': hp.uniform('subsample', 0.5, 1.),
    'reg_alpha': hp.uniform('reg_alpha', 0.01, 1.),
    'reg_lambda': hp.uniform('reg_lambda', 0.01, 1.),
}
```

running hyperopt

```
lgb_objective.i = 0
best = fmin(fn=lgb_objective,
            space=lgb_parameter_space,
            algo=tpe.suggest,
            max_evals=200)
```

2 SB ML Risk Model: *LIME*

- Explains the predictions of *any* classifier by learning an interpretable model locally around the prediction

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)$$

Where ξ is referred as "the explanation", $\mathcal{L}$ is measure of how unfaithful g is in approximating f in the locality defined by π_x and $\Omega(g)$ is measure of complexity

Example: classifying mushrooms as "*edible*" or "*poisonous*"

Tabular data

Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

http://savvastjortjoglou.com/intrepretable-machine-learning-nfl-combine.html

3 SB ML Risk Model Results

- **Test results:**

In an scenario where ~10% of the customers claim the f_1 score is:

f1_score(real, predicted) = 0.37
f1_score(real, predicted, average=‘*weighted*') = 0.84

	Claim Probability	*Severe Probability*	*Net Premium*
Marisa	0.81	0.08	76
Mark	0.83	0.25	709
Javier	0.81	0.31	161
Dani	**0.81**	**0.34**	**675**

!

4 The "*fight*" with the business

"Let's see if the algorithm works and then we see how we put it in production"
Managing Director - MGA

"Until I feel comfortable we won't use it"
Head of Underwriting

"We need to understand it to use it" ; "I *am not sure we need this*"
Senior Underwriter Analyst

Thanks!

Any questions ?

You can find me at:

- javier.rodriguez@simplybusiness.co.uk