Edit and Imputation of LCF survey data with Machine Learning

Claus Sthamer & David Sampson

Data Science Campus

1st April 2020

Claus.Sthamer@ons.gov.uk

Editing of LCF Income data with ML

Current process:

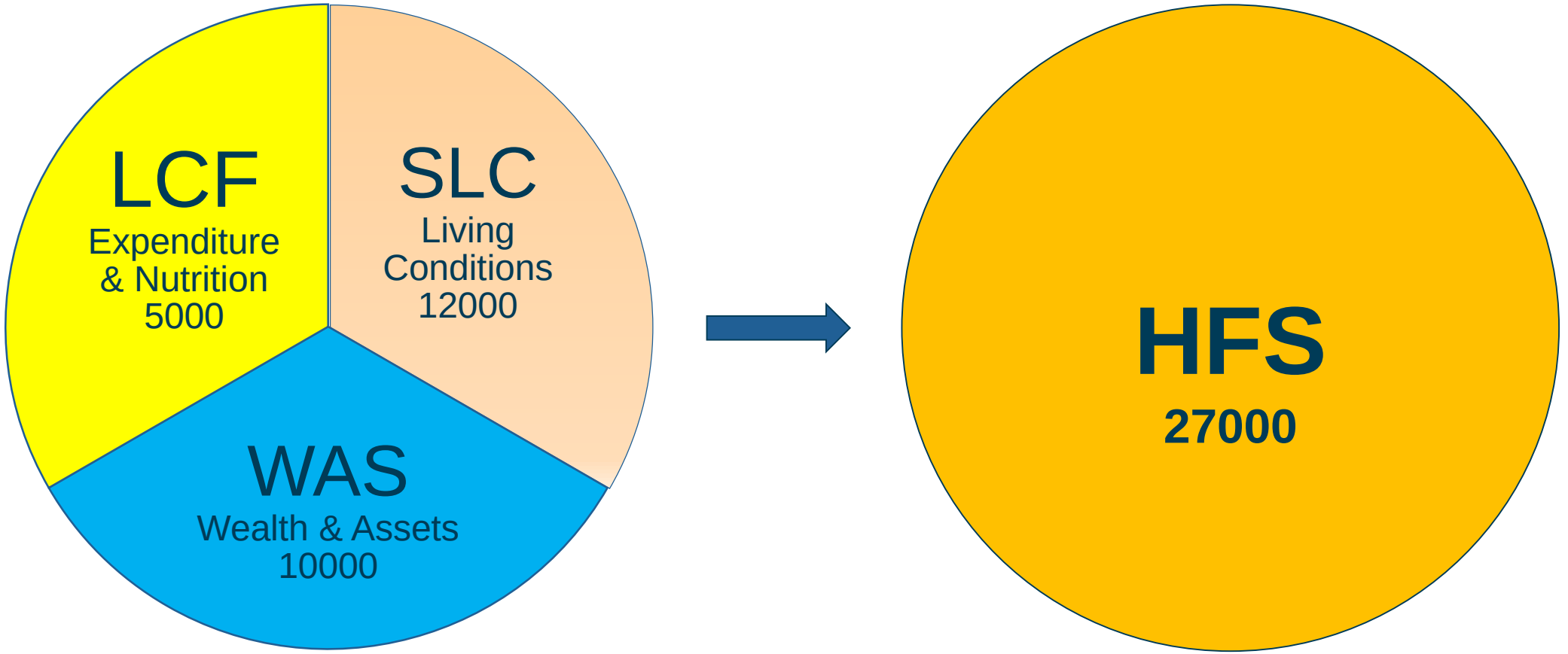
All LCF data are manually examined by the Editing team in Titchfield
Incorrect values are corrected and missing values (Don't Knows & Refusals) are inserted

ML:

Can ML make this more efficient?
Can ML predict enough records that need to be looked at?
Can ML reduce time spent on records that do not need to be worked on?

Claus.Sthamer@ONS.gov.uk

The HFS and it's component Surveys



Claus.Sthamer@ONS.gov.uk

Definitions:

Edit:
Identifying records that need values changed or missing values inserted

Imputation:
Changing and Inserting missing values

Claus.Sthamer@ONS.gov.uk

Survey Specific Editing (Currently in Production)

LCF – all cases go through clerical editing

too slow and too labour intensive → **Speed &Cost?**

SLC – Scripted outlier detection (range of values)

Only about 10% of changes that are made with the LCF method are made with the SLC method → **Accuracy?**

WAS – Scripted outlier detection (range of values)

Accuracy?

→ **ML**

Claus.Sthamer@ONS.gov.uk

Why a new solution for Editing?

LCF

- Can the LCF method be upscaled?
- Over Editing?

SLC

- Can the scripted outlier detection scripts be made more accurate?
- Can better rules be found than just value ranges?

Claus.Sthamer@ONS.gov.uk

ML Algorithm

Office for National Statistics

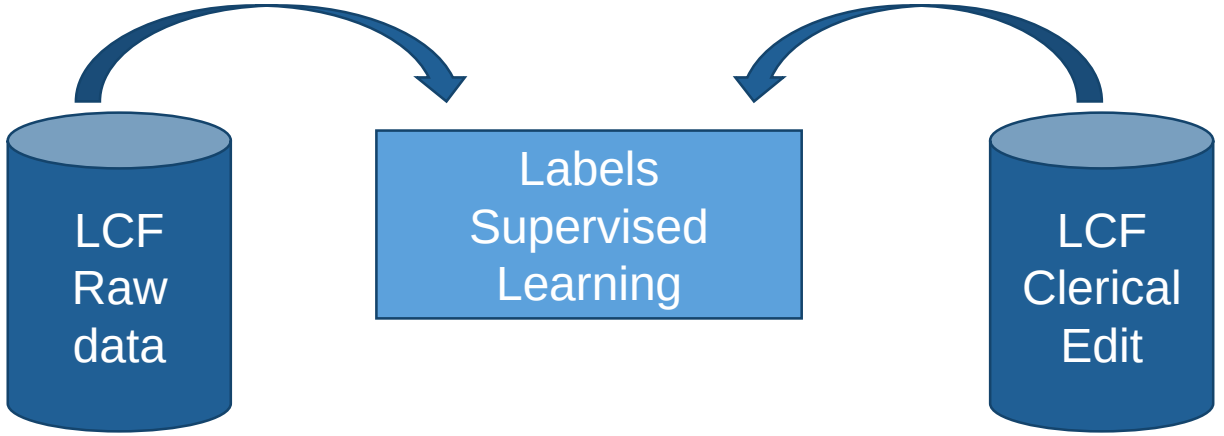
Claus.Sthamer@ONS.gov.uk

What do we need?

Random Forest: due to the small number of records

For LCF we have:

Training Data:	LCF 8Q3	3059 Records	(2018 quarter 3)
Test Data:	LCF 8Q2	2912 Records	(2018 quarter 2)



Claus.Sthamer@ONS.gov.uk

Why use 8Q2 LCF test data?

LCF 8Q2 raw data were put through the SLC scripted outlier detection method.

Only about 10% of the changes compared to the full clerical method were made.

SLC method might be inadequate?
LCF clerical method might be over editing?

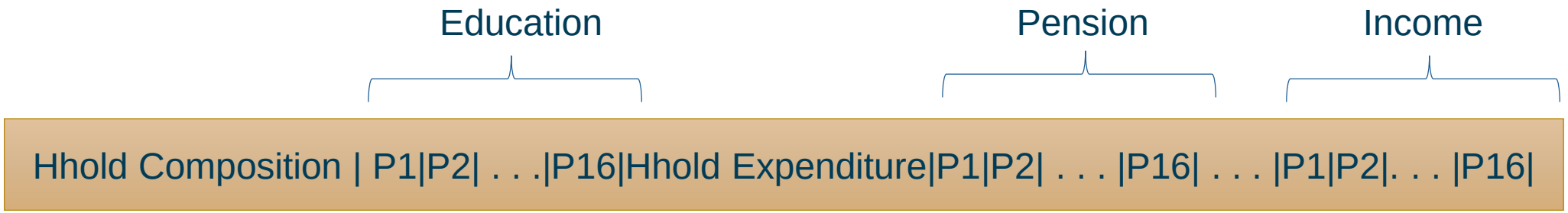
Claus.Sthamer@ONS.gov.uk

Data

Claus.Sthamer@ONS.gov.uk

LCF Household Record contains:

- All questionnaire data collected from one Household
- A household can have up to 16 persons (P1 …. P16)
- Various blocks with arrays for P1 to P16

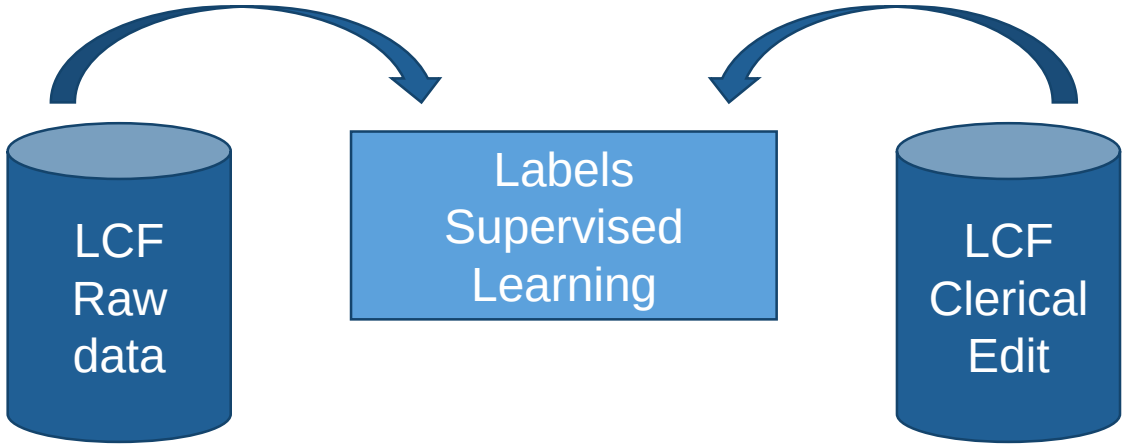


Area	Address	Hhold	Person	NetPay	IncTax
1201	2	1	1	3240	23
1201	2	1	2	1350	375

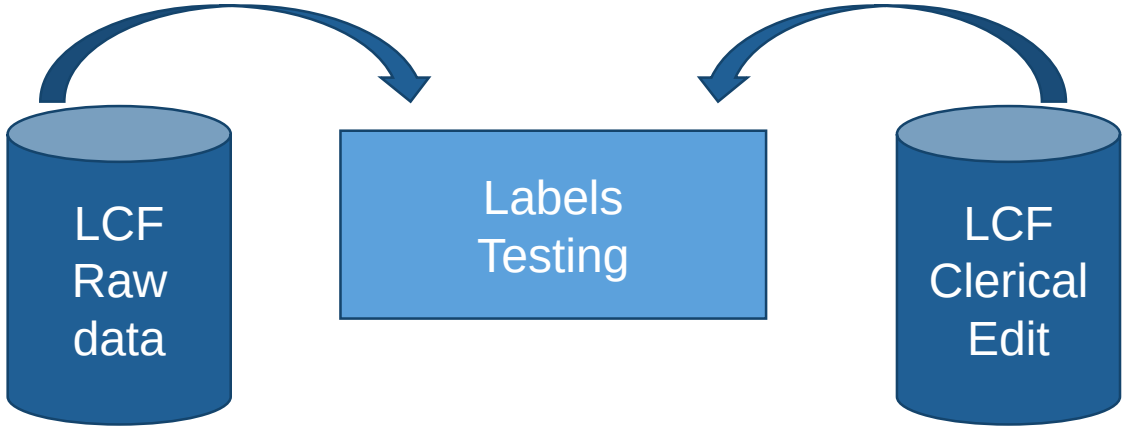
Claus.Sthamer@ONS.gov.uk

8Q2 & 8Q3 Person Level data files

8Q3



8Q3



Claus.Sthamer@ONS.gov.uk

Person level Data – Feature Selection

Of the about 2000 features (guess work) Features that describe a person

I selected 106 from these areas:

- Income
- Education
- Family situation
- Job and secondary job details
- Income and tax of job and secondary job
- Happiness and wellness
- Affordability of hobbies, clothes and shoes

Claus.Sthamer@ONS.gov.uk

Change Vector Features & Frequency for 8Q2

Selected from "2018 Income LCF Editing Instructions.docx"

Main Job		
NetPay	-Last take home pay including	81
NetPd	-Period coverd for NetPay	48
Upd	-How often are you usually paid	18
IncTax	-Income Tax	235
Taxref	-Tax Refund? Yes/No	39
NIns	-National Insurance Contribution	254
GrossPay	-Last gross pay from main job	336
GrossPd	-Period coverd for last Gross Pay	68
GrossTel	-Total personel Income before Tax	632
UNett	-Amount usually received AFTER all deductions	10
UGross	-Amount usually received BEFORE all deductions	19
Bonus	-Number of bonuses received in last12 months	6
BonAm	-Amount of bonus	8
DedPenAm	-Deduction for pension or superannuation	113
SJbGrs	-Last gross pay from main job	8
RedAmt	-Redundancy payment in last 12 months	0
Benefit		
ComBAm	-Combined Benefit Amount	34

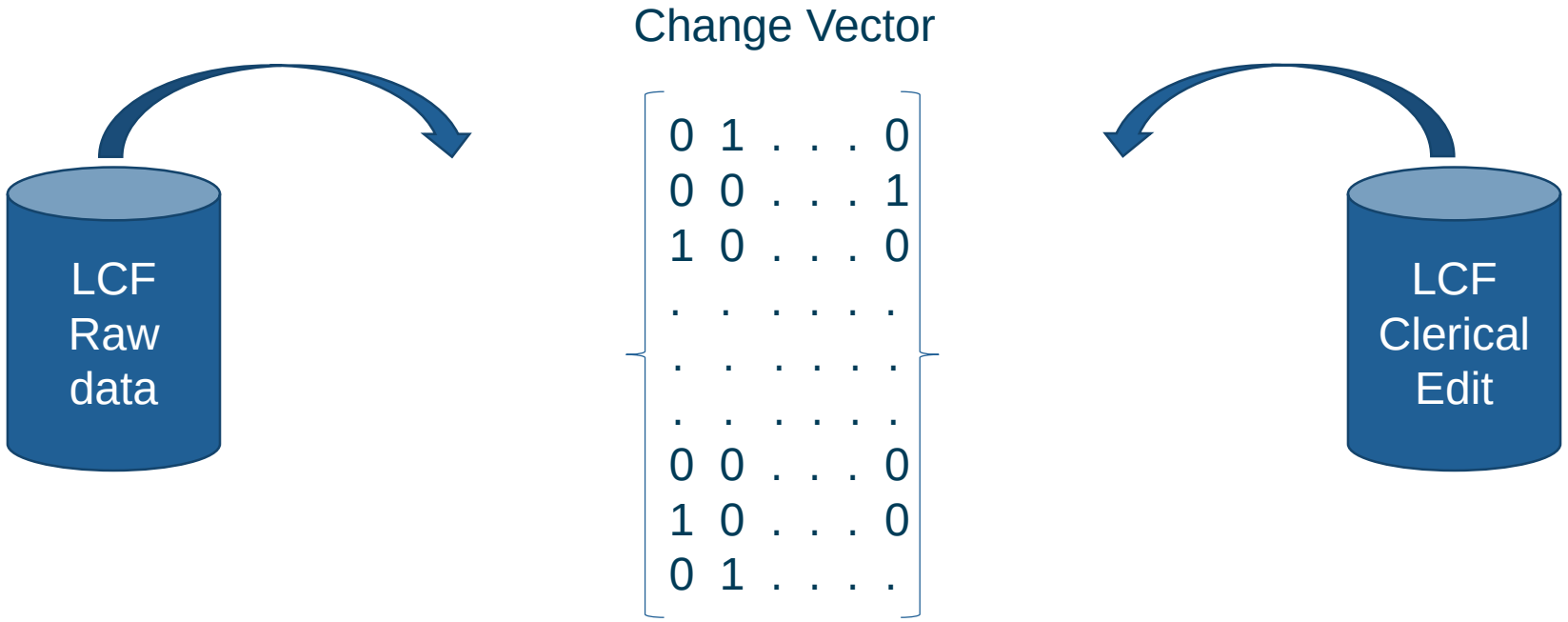
Subsidiary job as self-employed		
Profit1	-profit/loss amount	8
PrBefore	-Amount of profit BEFORE Tax	8
TaxDAmt	-Amount of income tax deducted last time	0
NIDAmt	-NI deducted last time	1
OwnAmt	-Last 12 months average take each month	1
SelfEmployment		
SEInc	-Average Income after paying costs	10
SeNIRAmt	-Last NI payment	51
SeTaxAmt	-Total tax payed last 12 months	6

Claus.Sthamer@ONS.gov.uk

Calculate the Change Vectors:

8Q3 - Training

8Q2 - Testing

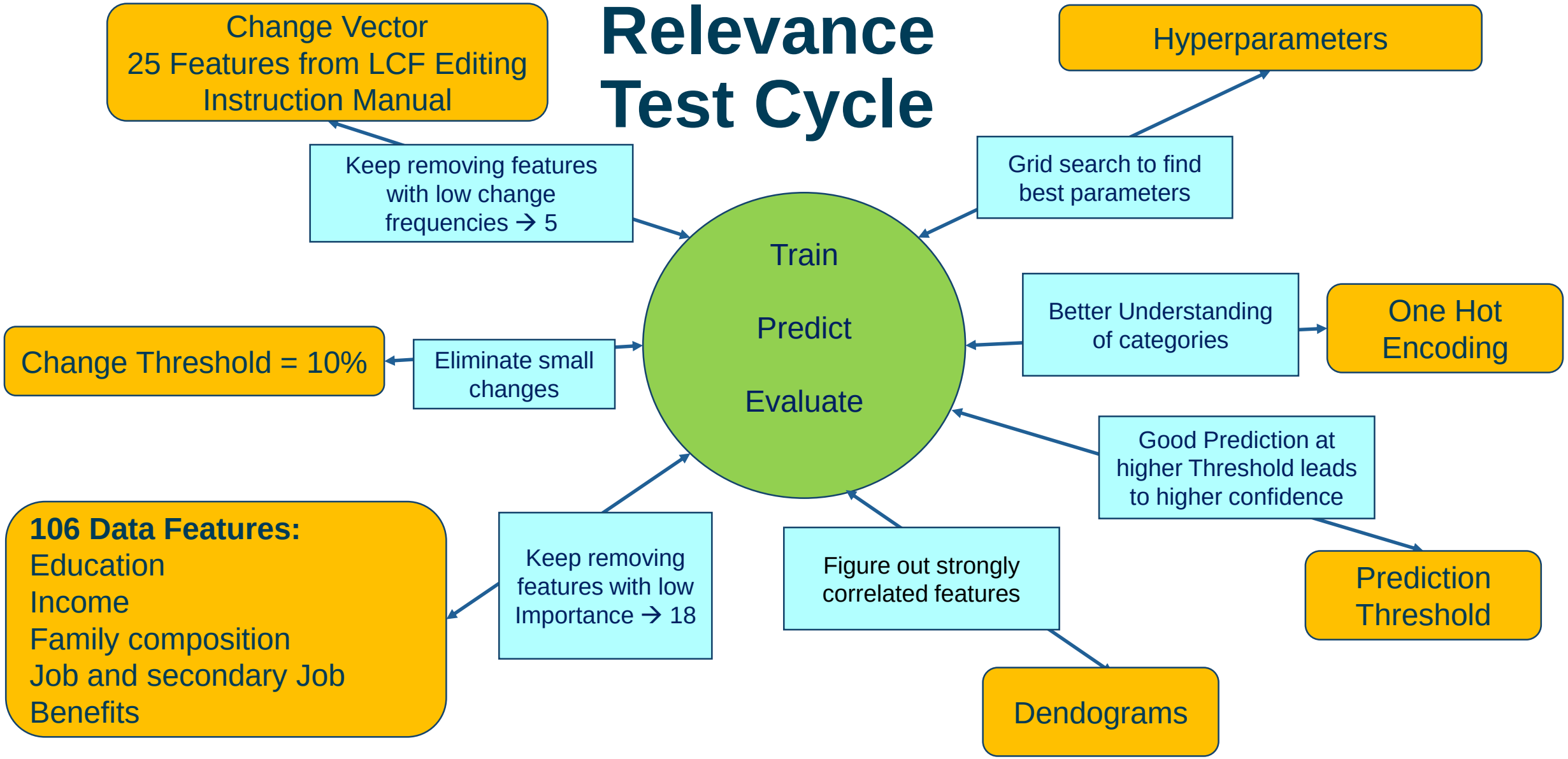


0 indicates No-Change
1 indicates a Change the Features

8Q3 - Records labelled as Change: 442 out of 3059 records
8Q2 - Records labelled as Change: 451 out of 2912 records

Claus.Sthamer@ONS.gov.uk

Relevance Test Cycle



Claus.Sthamer@ONS.gov.uk

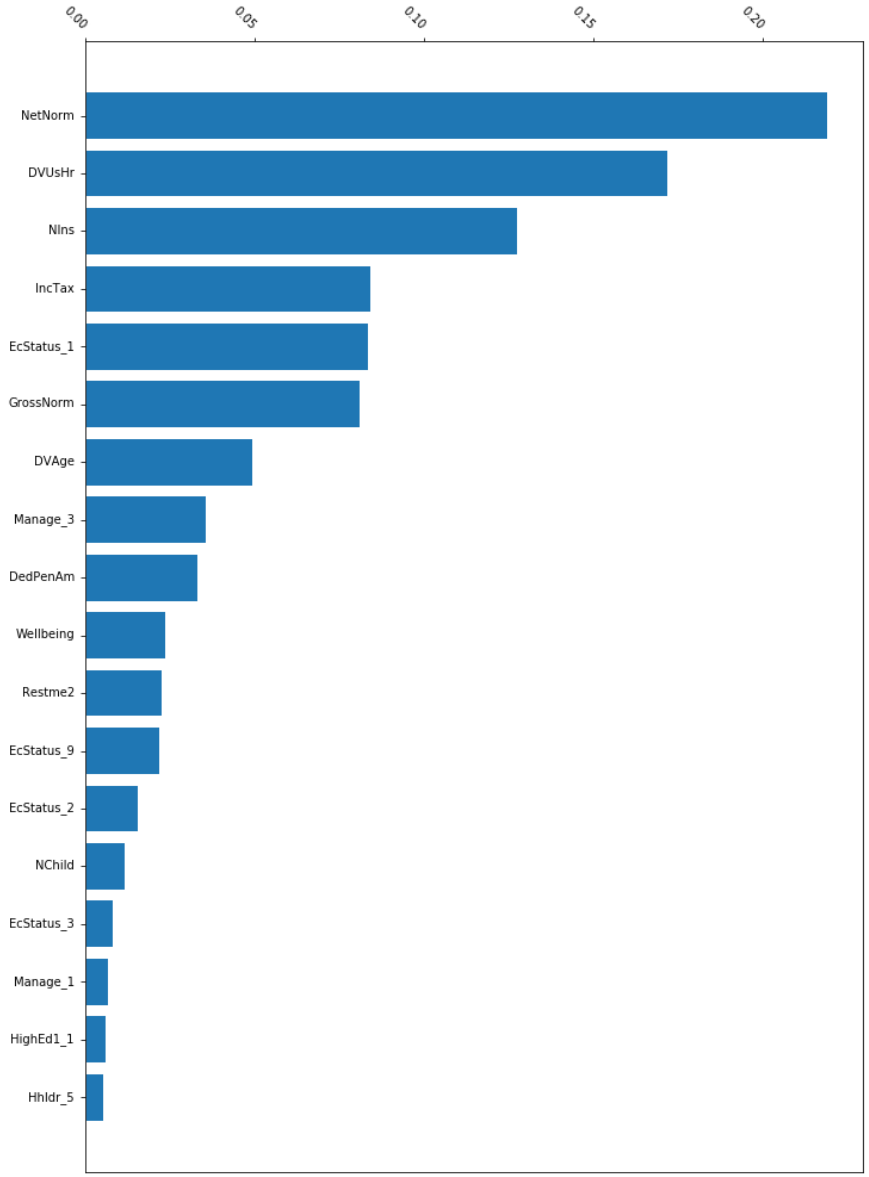
Feature Importance

NetPay	0.1433196937971856 8
DVAge	0.0500789754305877 99
DVUsHr	0.0498281114243205 18
GrossTel	0.0481397578301529 03
GrossPay	0.0445478489096711 87
Restme2	0.0433326558683784 92
.	
.	
.	
StartJ	0.0014887652098582085
NCUnd5	0.0013304446062182992
Solo	0.0008001098357992231 5

Claus.Sthamer@ONS.gov.uk

Features

Name	Description
NetNorm	Annual amount of net income (after deductions)
DVUsHr	Derived number of hours worked per month
NIns	National Insurance paid over given period
IncTax	Income Tax paid over given period
EcStatus == 1	Person works full time
GrossNorm	Annual amount of gross pay (before deductions)
DVAge	Derived age of person
Manage == 3	Person doesn't manage anybody
DedPenAm	How much pension has been paid over given period
Wellbeing	Aggregate of four quality of life features (Satisfaction, Worth, Happy, Anxiety)
RestMe2	How many years has the person resided at address
EcStatus == 9	Person is retired
EcStatus == 2	Person works part time
NChild	Number of children the person has
EcStatus == 3	Person works full time and is self employed
Manage == 1	Person manages someone/people
HighEd1 == 1	Highest qualification this person has achieved is a degree
Hhldr == 5	The person is not the named owner/renter of address



David.Sampson@ONS.gov.uk

Train & Test:

```
# Initialise the Random Forest
net_pay_Tree = ensemble.RandomForestClassifier(bootstrap = True,
                                               class_weight = 'balanced_subsample',
                                               criterion = 'gini',
                                               max_depth = 11,
                                               max_features = 'sqrt',
                                               max_leaf_nodes = 230,
                                               min_samples_leaf = 12,
                                               n_estimators = 180,
                                               n_jobs = -1 )
# Training the Random Forest
net_pay_Tree = net_pay_Tree.fit(df_pre_edit_8Q3,df_change_8Q3)

# Predicting for 8Q2 data the need to change any of the 5 features
list_test_proba = net_pay_Tree.predict_proba(df_pre_edit_8Q2)

# Prediction Result
NetPay      [[0.81181041329, 0.18818958671], [0.8837823393...
IncTax      [[0.720982389396, 0.279017610604], [0.80501230...
NIns        [[0.633917438321, 0.366082561679], [0.77009904...
GrossPay    [[0.608604644881, 0.391395355119], [0.80924280...
DedPenAm    [[0.807707769904, 0.192292230096], [0.93019158...
```


Claus.Sthamer@ONS.gov.uk

Results

Change vector with number of changes:

NetPay	-Last take home pay including	81
IncTax	-Income Tax	235
NIns	-National Insurance Contribution	254
GrossPay	-Last gross pay from main job	336
DedPenAm	-Deduction for pension or superannuation	113

Training Data: 8Q3 3059 Person records, 442 labelled as Change, reduced to 362 with 10% change threshold
Test Data: 8Q2 2912 Person records, 451 labelled as Change , reduced to 361 with 10% change threshold

Features:

106 features reduced to 18 by removing features with low Importance

Prediction Threshold:	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**
Recall	97.2%	94.6%	88.1%	84.0%	78.6	75.2	72.9
Precision	35.7%	40.0%	45.4%	54.5%	67.1	81.3	90.1
F1-Score	52.2%	56.2%	59.1%	66.1%	72.4	78.1	80.6
TP	376	366	341	325	304	291	282
FP	677	550	410	271	149	67	31

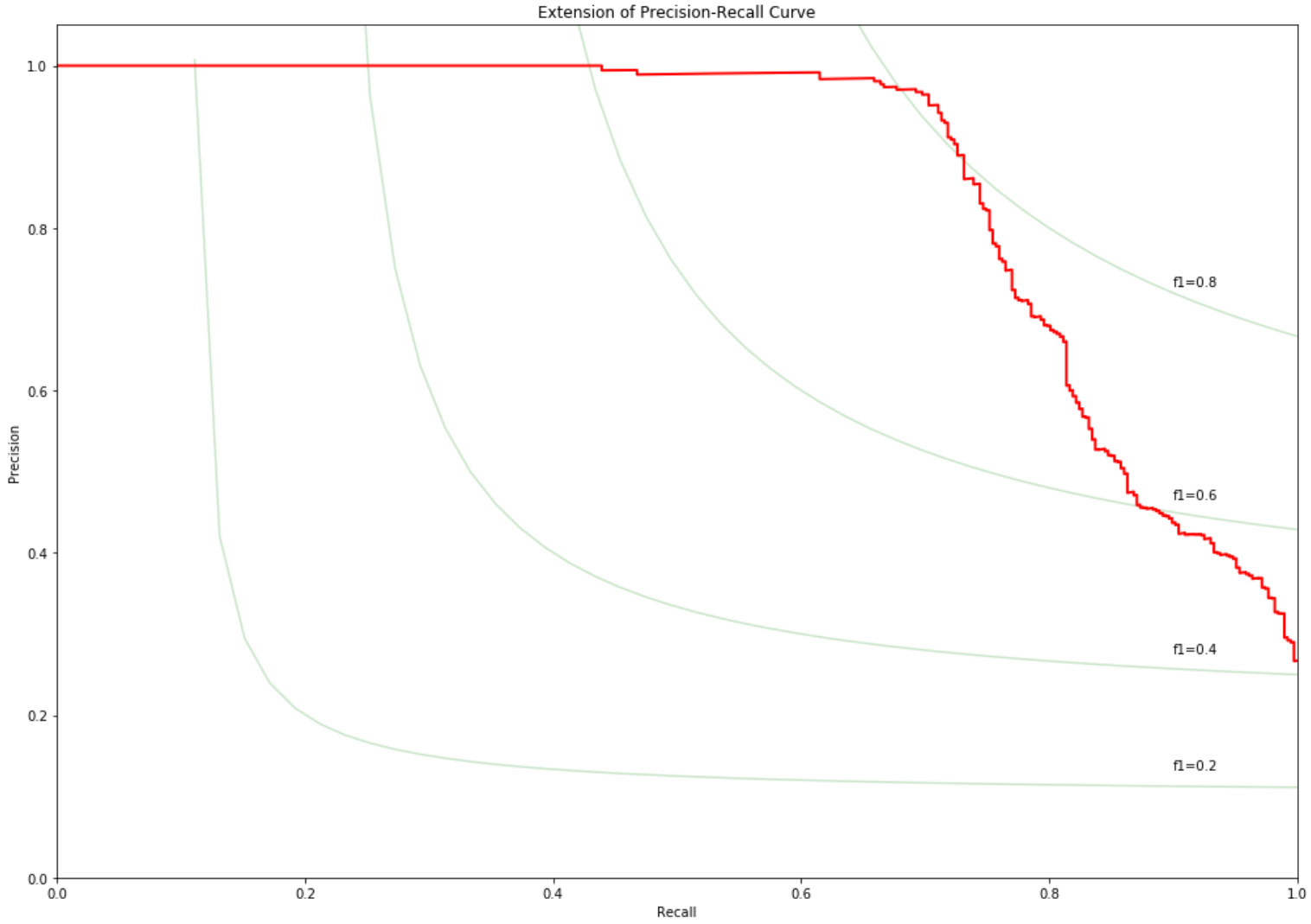
Claus.Sthamer@ONS.gov.uk

Prediction Confidence

Cases with at least one predicted change, the darker the colour the higher the higher the confidence:

	Area	Address	HHold	Person	NetPay	IncTax	NIns	GrossPay	DedPenAm
1980					2849	591.27	377.76	4218	337
420									
1660					1600	220	155	2100	125
2391					1010	30	27	1060	
1967									
1019					1200	300	67		53
1674					370	60	25	450	
838					1200	240	162	1602	
2818					1300	101.99	95.45	1497.44	

Claus.Sthamer@ONS.gov.uk



David.Sampson@ONS.gov.uk

Explainability

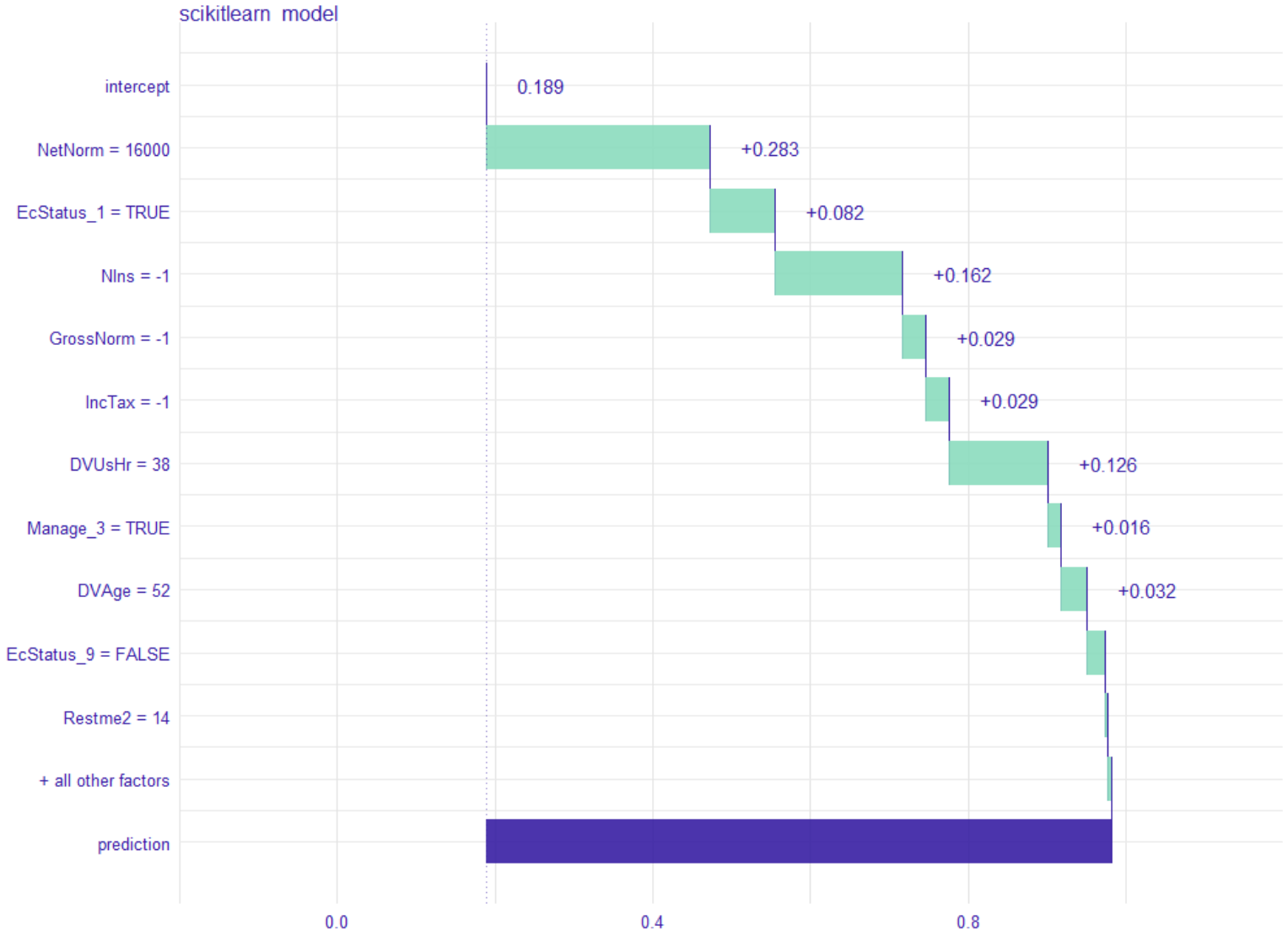
David.Sampson@ONS.gov.uk

Explainability

- Why?
  - Deeper dive for incorrect predictions
  - ICO & The Turing's guidance rooted in GDPR
    - "Project expl***AI***n"
    - ***"Be transparent****: make your use of AI for decision-making obvious and appropriately explain the decisions you make to individuals in a meaningful way."*
- How?
  - Tools in Scikit-learn; R packages such as iBreakDown & DALEX

https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2019/12/ico-and-the-alan-turing-institute-open-consultation-on-first-piece-of-ai-guidance

David.Sampson@ONS.gov.uk

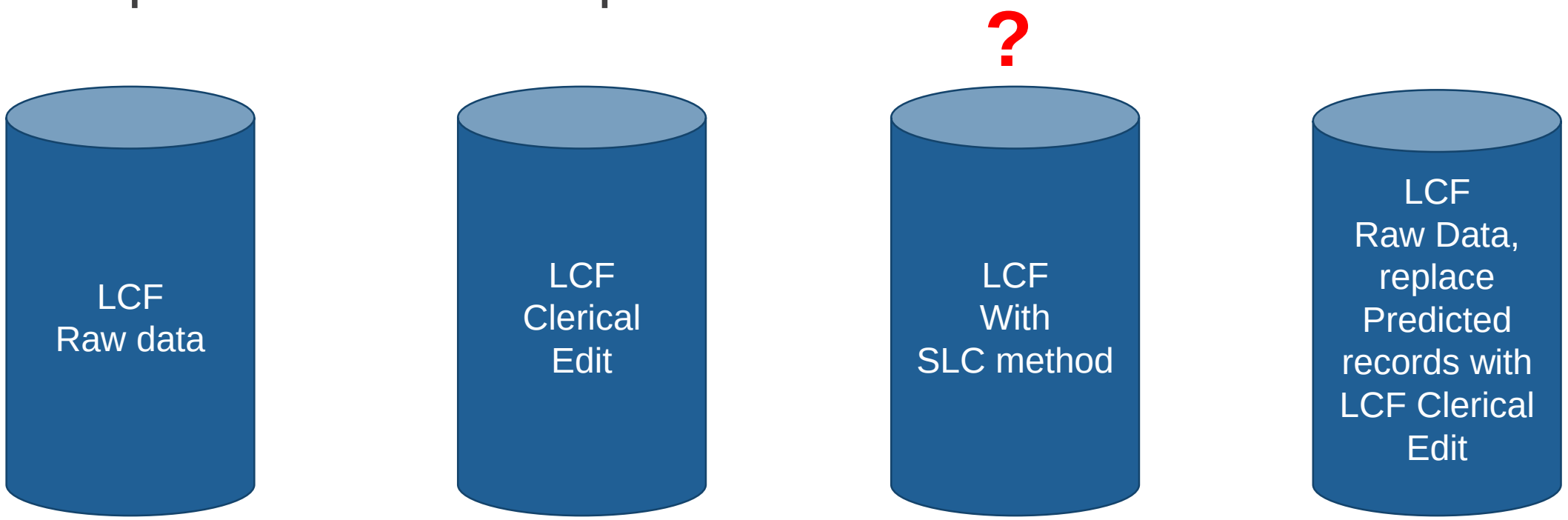

scikitlearn model
intercept
NetNorm = 16000
EcStatus_1 = TRUE
NIns = -1
GrossNorm = -1
IncTax = -1
DVUsHr = 38
Manage_3 = TRUE
DVAge = 52
EcStatus_9 = FALSE
Restme2 = 14
+ all other factors
prediction
0.189
+0.283
+0.082
+0.162
+0.029
+0.029
+0.126
+0.016
+0.032
0.0
0.4
0.8

What is next?

Claus.Sthamer@ONS.gov.uk

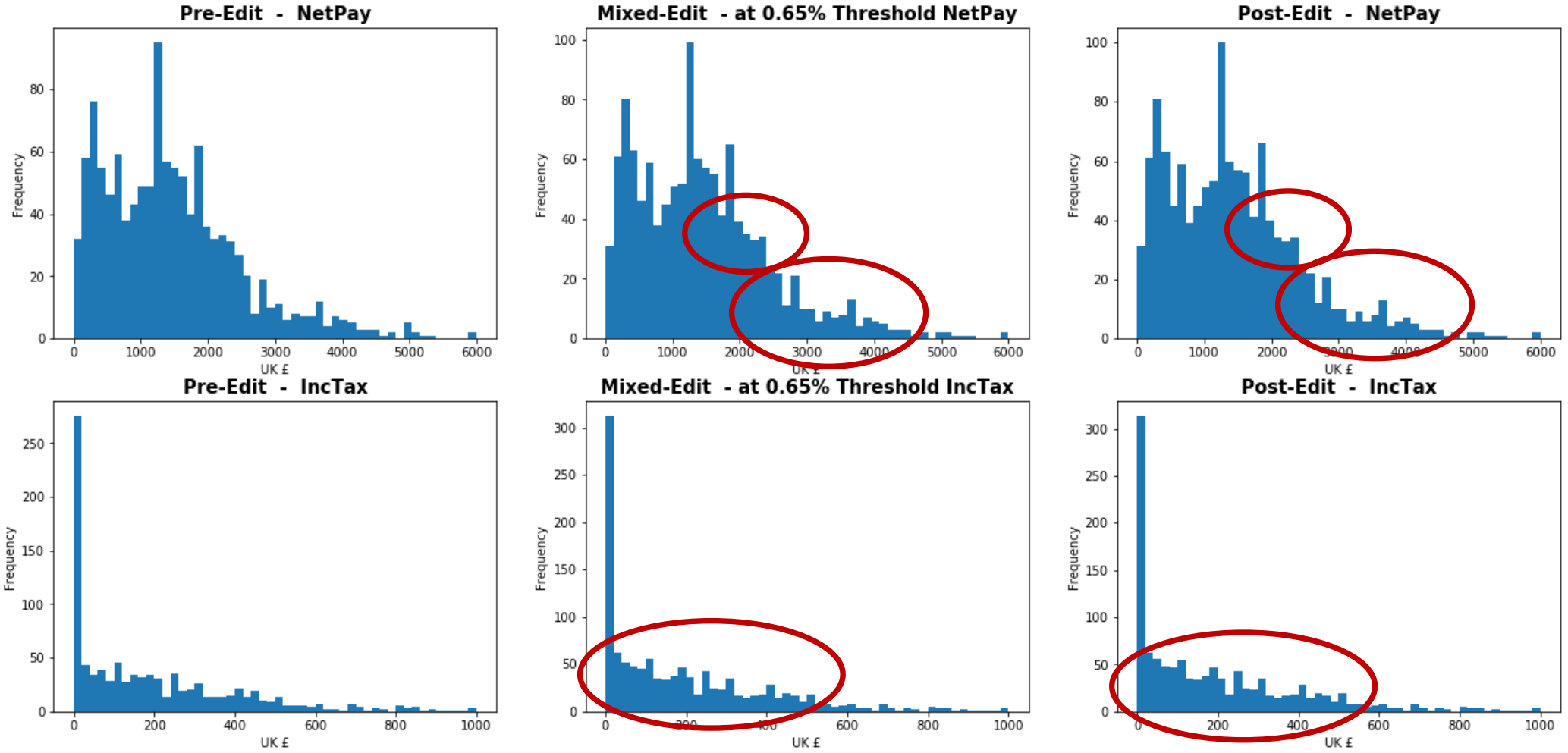
Next Steps

Compare these 4 Outputs:



Claus.Sthamer@ONS.gov.uk

Is the ML solution good enough?



Claus.Sthamer@ONS.gov.uk

Prototype?

To assist the Clerical Editing of LCF data in Titchfield

Legacy  Transformation?

Claus.Sthamer@ONS.gov.uk

Cloudera FFL Workshop

ONS's broader engagement with our technology partner, Cloudera

Two day workshop led by Cloudera's Data Science & Machine Learning-focused subsidiary, Fast Forward Labs (FFL).

During the workshop CFFL helped ONS examine its approach to using Machine Learning in its production of statistics, The strategic review will consider three aspects of using machine learning in the way described above:

- Accuracy
- Explainability
- Ease of integration into existing business processes

Workshop 9 &10 March ONS Newport

Eric.Deeben@ONS.gov.uk