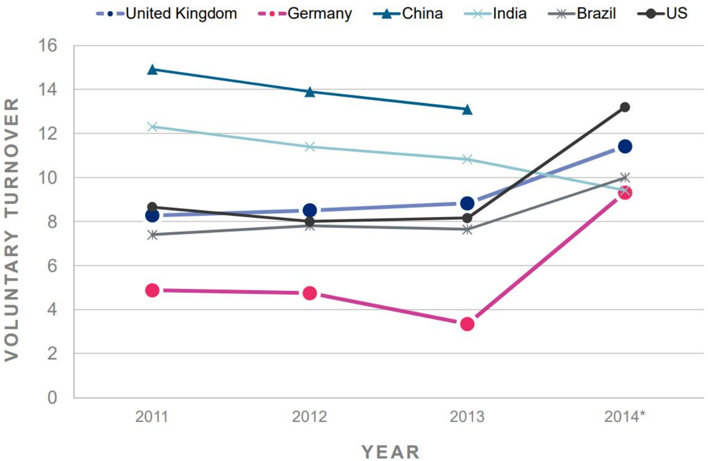
Resignation Case Study

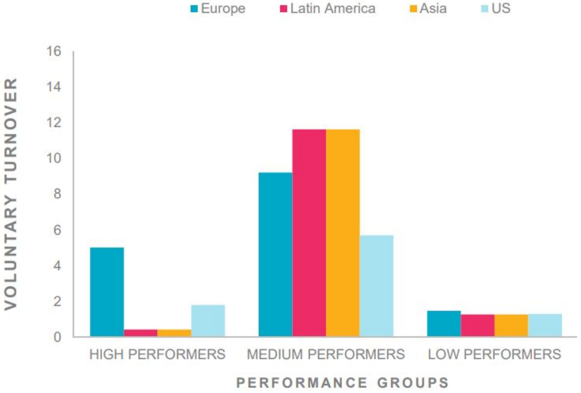
Shilpa

Business problem

- Company ABC noted **16%** high performer attrition. This is more than ~**2x** market voluntary turn
- **Goal** - To construct a resignation prediction model to enable the business reduce their regrettable attrition

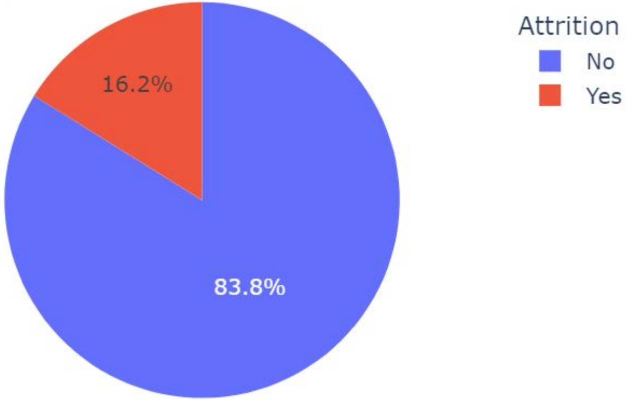
Market vs Internal





Company ABC

fig 1. attrition data distribution



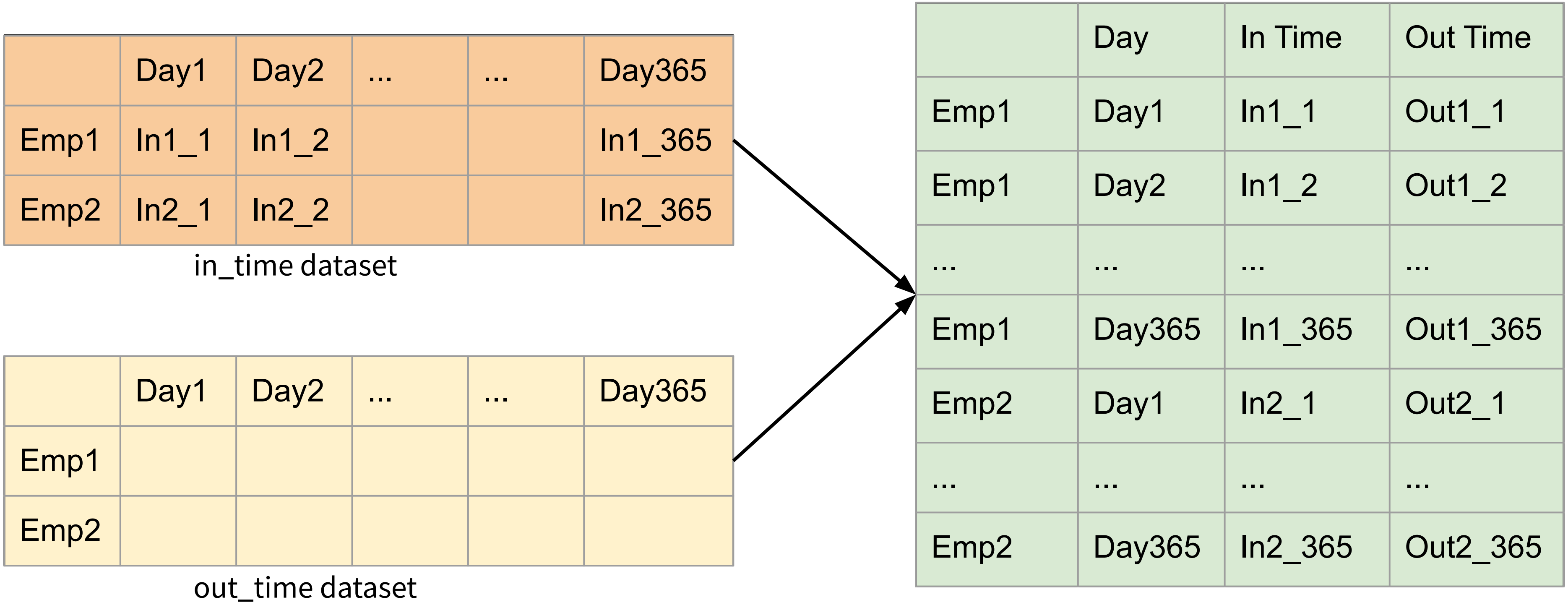
Source: Mercer
‘Trends and drivers of workforce turnover survey’ 2015

Datasets

File name	Description
data_dictionary	Definition of all variables available for study
employee_survey_data	Employee survey inputs
manager_survey_data	Manager survey inputs
general_data	Employee descriptive variables and attrition
in_time	Employee clock-in times
out_time	Employee clock-out times

Feature Engineering

Preliminary EDA + Feature engineering *(1 of 4)*



	Day1	Day2	...	...	Day365
Emp1	In1_1	In1_2			In1_365
Emp2	In2_1	In2_2			In2_365

in_time dataset

	Day1	Day2	...	...	Day365
Emp1					
Emp2					

out_time dataset

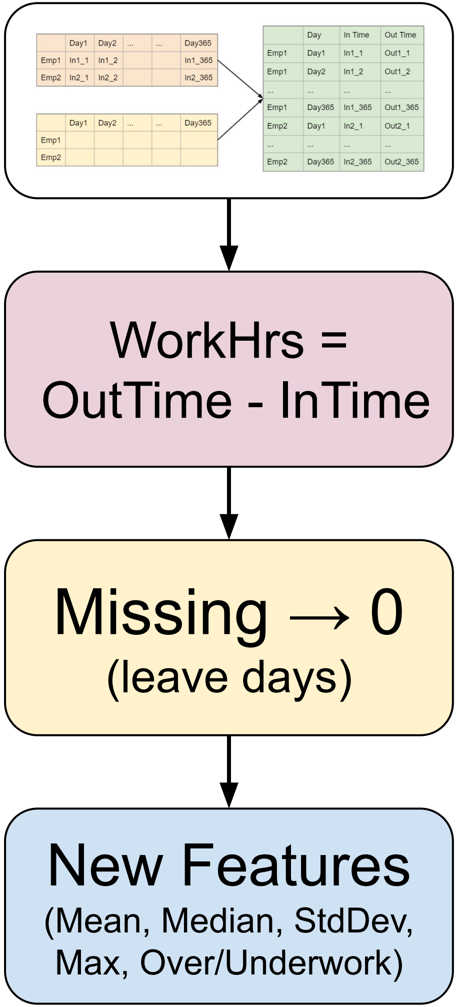
	Day	In Time	Out Time
Emp1	Day1	In1_1	Out1_1
Emp1	Day2	In1_2	Out1_2
...	...	...	...
Emp1	Day365	In1_365	Out1_365
Emp2	Day1	In2_1	Out2_1
...	...	...	...
Emp2	Day365	In2_365	Out2_365

Raw

fig 2. Time datasets transformation

Transformed

Preliminary EDA + Feature engineering *(2 of 4)*



	Day1	Day2	...	...	Day365
Emp1	In1_1	In1_2			In1_365
Emp2	In2_1	In2_2			In2_365

	Day1	Day2	...	...	Day365
Emp1					
Emp2					

	Day	In Time	Out Time
Emp1	Day1	In1_1	Out1_1
Emp1	Day2	In1_2	Out1_2
...	...	...	...
Emp1	Day365	In1_365	Out1_365
Emp2	Day1	In2_1	Out2_1
...	...	...	...
Emp2	Day365	In2_365	Out2_365

WorkHrs = OutTime - InTime

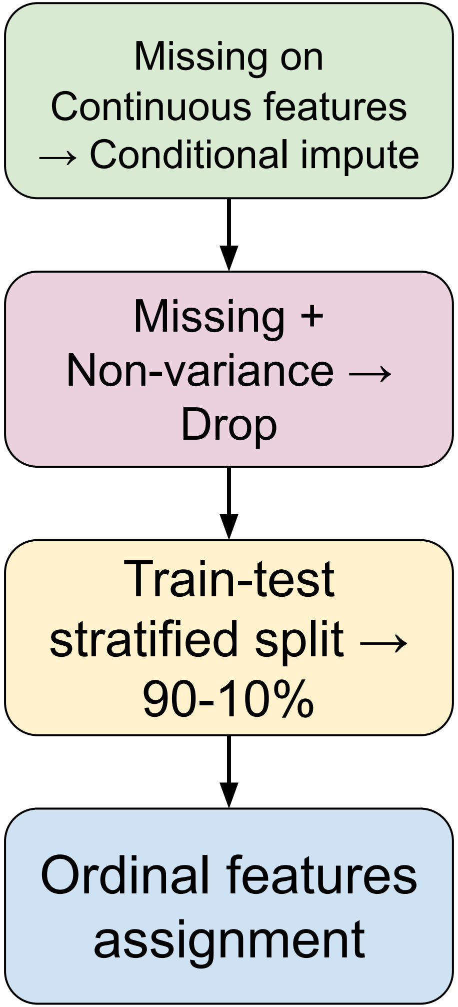
Missing → 0 (leave days)

New Features (Mean, Median, StdDev, Max, Over/Underwork)

Findings

1. Wide dataset layout with recorded work times per workday per employee for 2015
2. Missing data in dataset corresponding to public holidays or employee leaves
3. Standard hours in general_dataset can be compared with average actual employee work time to signal over or under work

Preliminary EDA + Feature engineering *(3 of 4)*



Missing on Continuous features → Conditional impute

Missing + Non-variance → Drop

Train-test stratified split → 90-10%

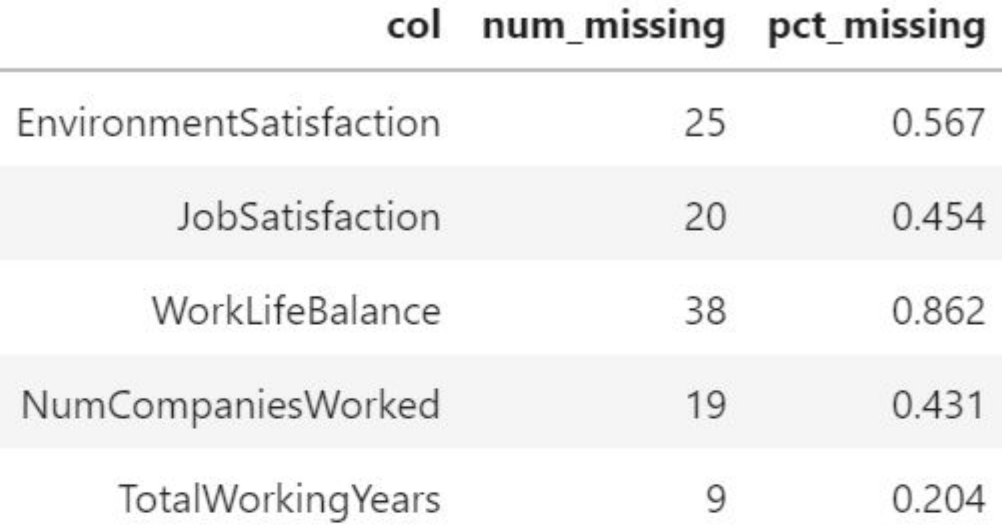
Ordinal features assignment

Findings

1. Missing feature - RelationshipSatisfaction, compared to data_dictionary
2. Features with missing values *(fig 3)* attributed to <1% of data distribution
    a. Ordinal + Continuous features
3. Non-variance features - EmployeeCount, Over18, StandardHours *(fig 4)*
4. Imbalanced distribution on response variable, attrition *(fig 1)*
5. Ordinal features - Education & all features in survey datasets

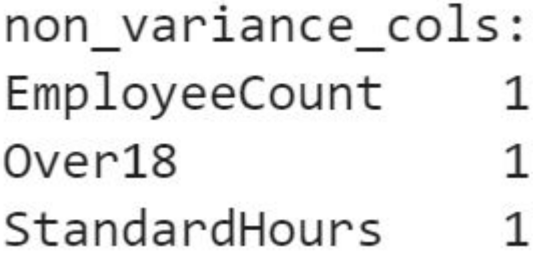
Preliminary EDA + Feature engineering *(4 of 4)*

fig 3. Features with missing



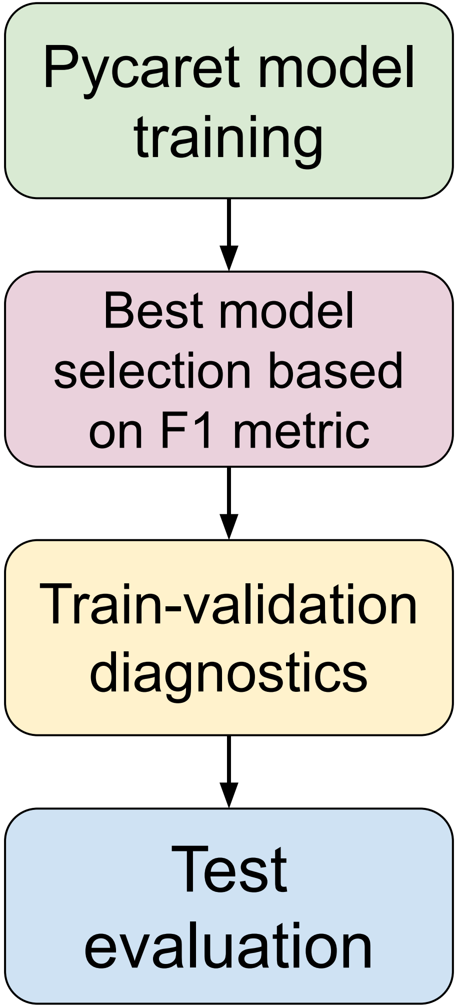
col	num_missing	pct_missing
EnvironmentSatisfaction	25	0.567
JobSatisfaction	20	0.454
WorkLifeBalance	38	0.862
NumCompaniesWorked	19	0.431
TotalWorkingYears	9	0.204

fig 4. Non-variance features



```
non_variance_cols:
EmployeeCount     1
Over18            1
StandardHours     1
```

Baseline Model



Pycaret model training

Best model selection based on F1 metric

Train-validation diagnostics

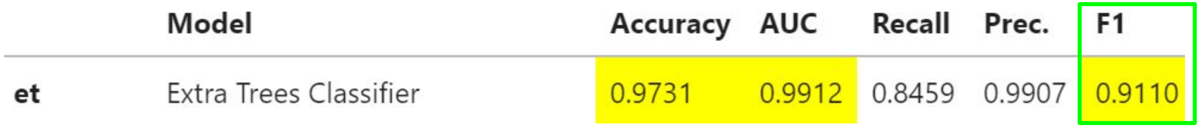
Test evaluation

Baseline resignation prediction model *(1 of 3)*

Metric	Split	Value	Fig
Accuracy	Train (10 folds)	0.97	5.1
	Test (holdout)	0.96	5.2
F1 (Since response is imbalanced)	Train (10 folds)	0.911	5.1
	Test (holdout)	0.86	5.2
Precision (Since response is imbalanced)	Train (10 folds)	1.0	5.4
	Test (holdout)	1.0	
Optimum probability threshold	Train	0.53	5.3

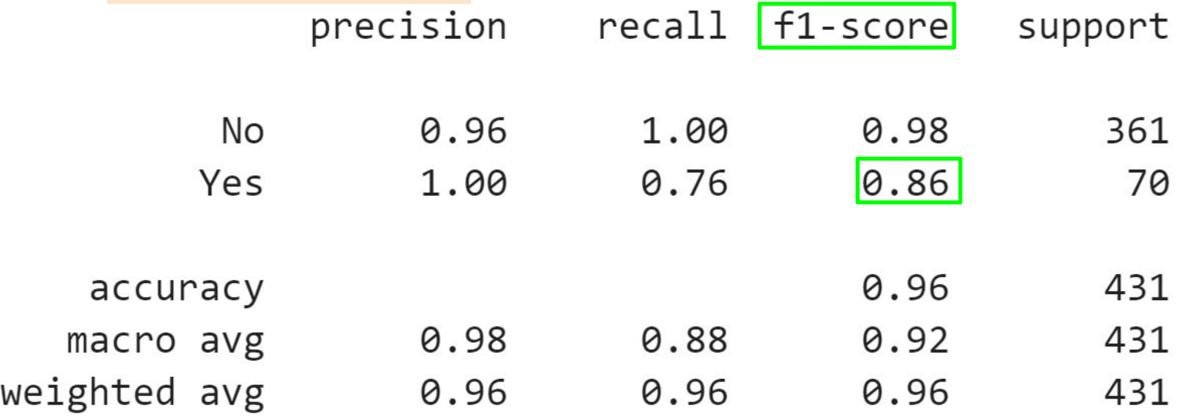
Baseline resignation prediction model *(2 of 3)*

fig 5.1. Best model avg 10-fold CV metrics



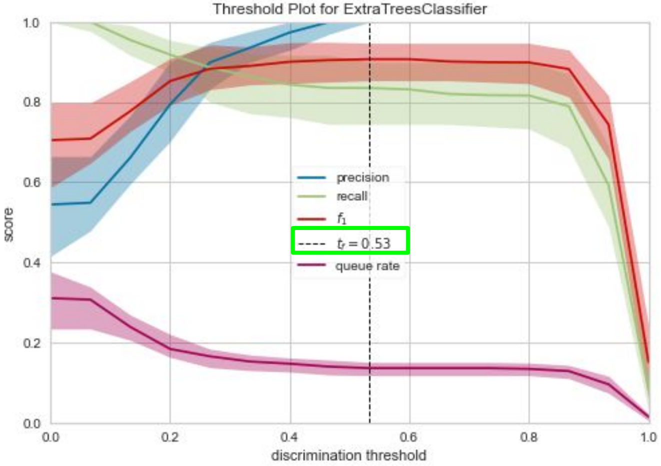
	Model	Accuracy	AUC	Recall	Prec.	F1
et	Extra Trees Classifier	0.9731	0.9912	0.8459	0.9907	0.9110

fig 5.2. Test metrics



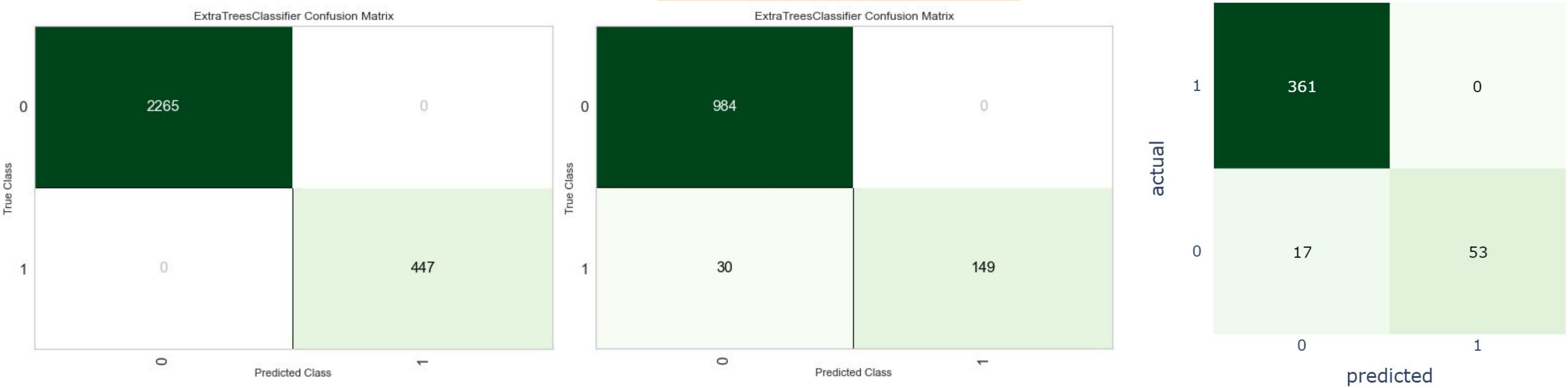
```
              precision    recall  f1-score   support

          No       0.96      1.00      0.98       361
         Yes       1.00      0.76      0.86        70

    accuracy                           0.96       431
   macro avg       0.98      0.88      0.92       431
weighted avg       0.96      0.96      0.96       431
```

fig 5.3. Model optimal threshold



Baseline resignation prediction model *(3 of 3)*

fig 5.4. Confusion matrix



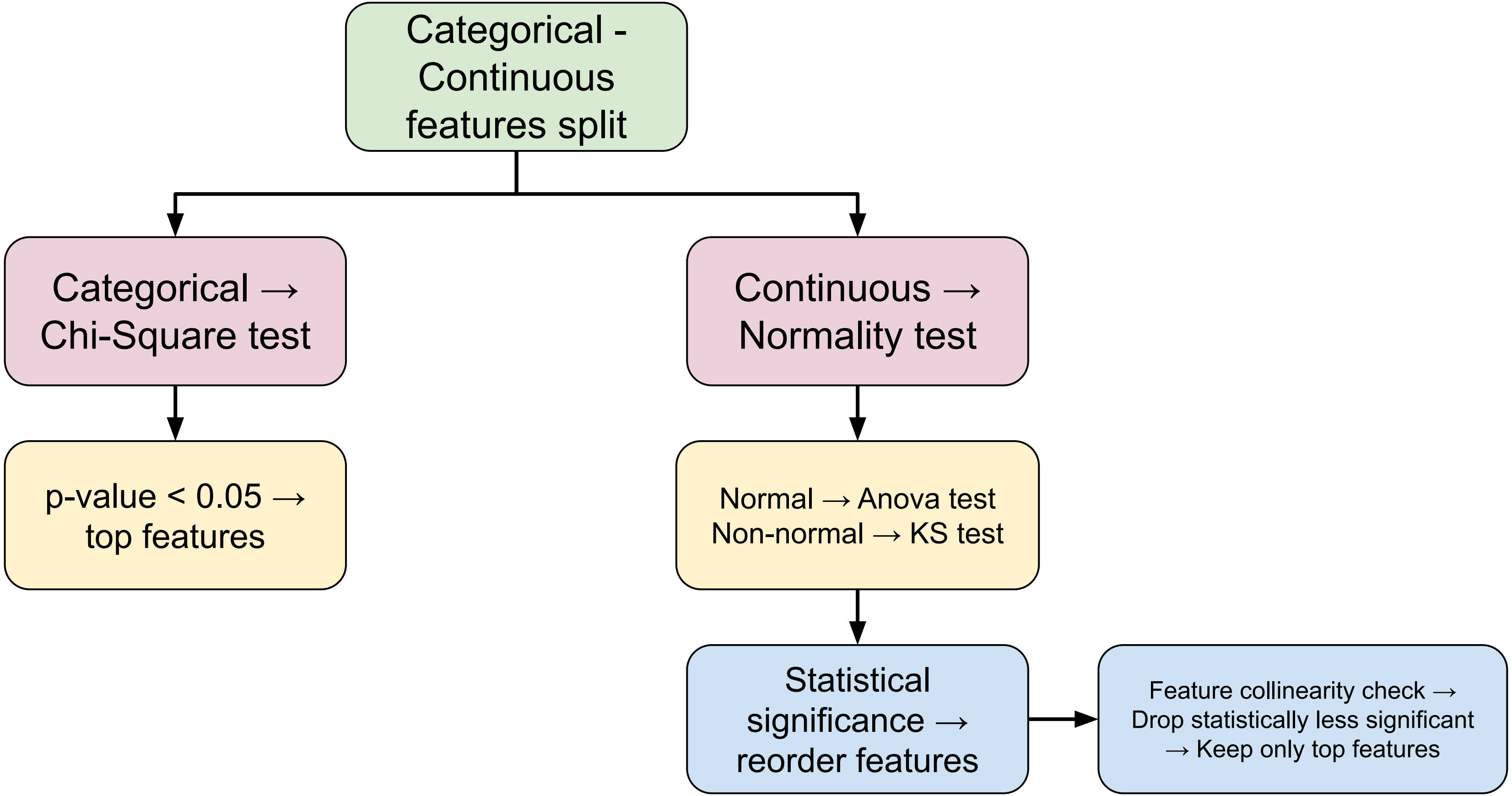
training

validation

testing

Feature Selection

Strategy



Categorical - Continuous features split

Categorical → Chi-Square test

p-value < 0.05 → top features

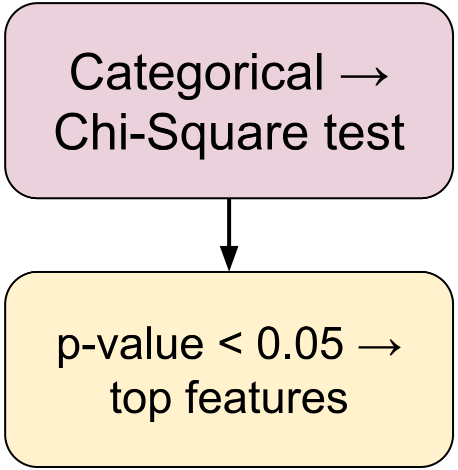
Continuous → Normality test

Normal → Anova test
Non-normal → KS test

Statistical significance → reorder features

Feature collinearity check → Drop statistically less significant → Keep only top features

Statistical feature selection EDA *(1 of 2): Categorical*



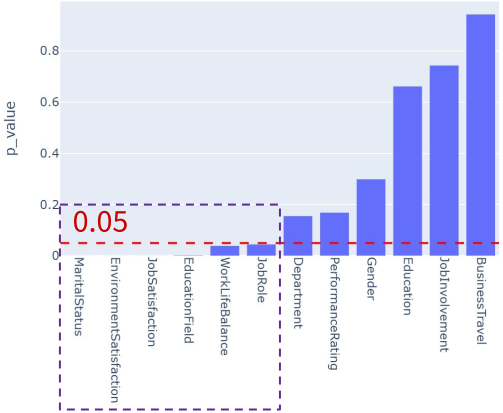
Categorical → Chi-Square test

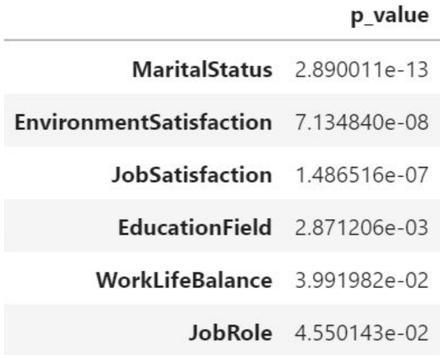
p-value < 0.05 → top features

Feature selection

1. Top features highly affecting response - MaritalStatus, EnvironmentSatisfaction, JobSatisfaction, EducationField, WorkLifeBalance & JobRole based on p-value <0.05 *(fig 6)*
2. Attrition insights for above features *(fig 7)*:
   a. High% resignees were single, scored poorly in employee survey, had a HR education background and worked as a Research Director

fig 6. Categorical feature importance

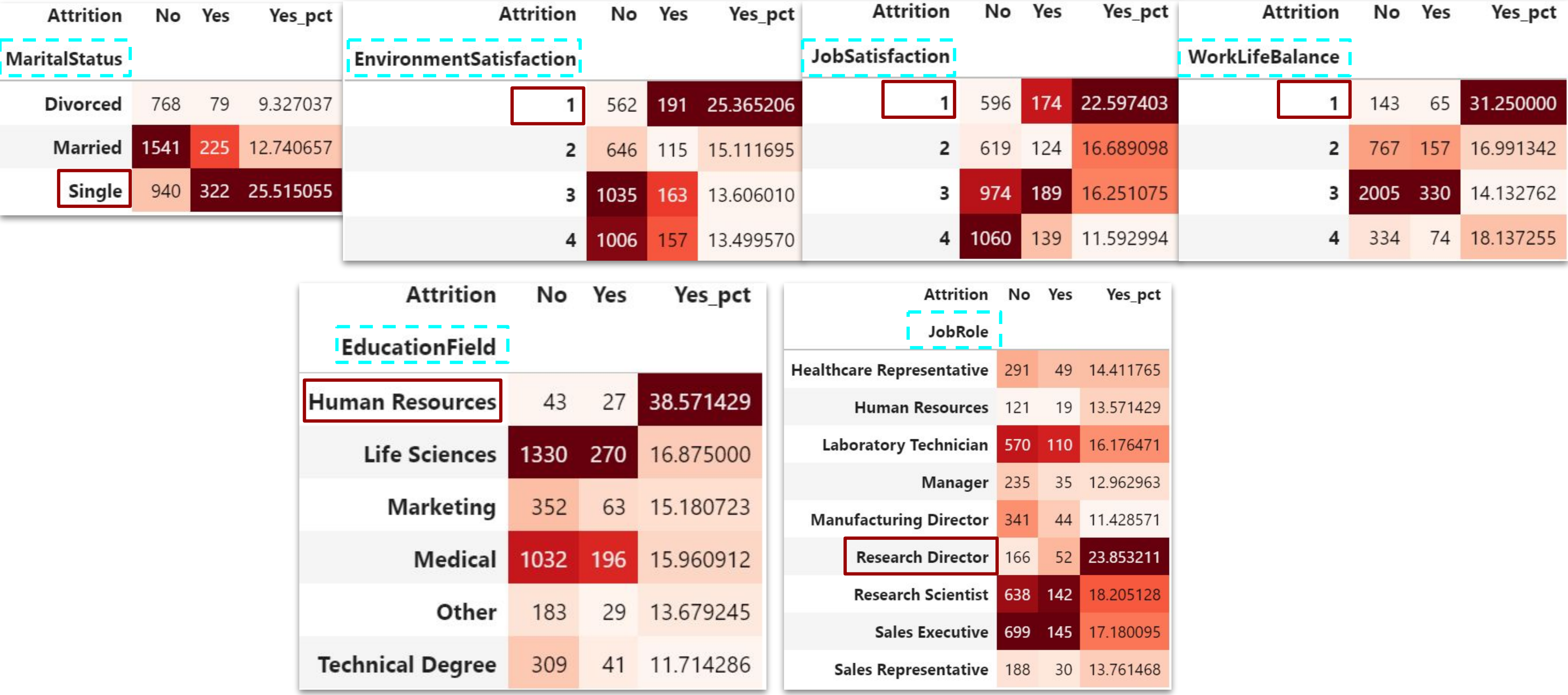




	p_value
MaritalStatus	2.890011e-13
EnvironmentSatisfaction	7.134840e-08
JobSatisfaction	1.486516e-07
EducationField	2.871206e-03
WorkLifeBalance	3.991982e-02
JobRole	4.550143e-02

Statistical feature selection EDA *(2 of 2): Categorical*

fig 7. Attrition insights - top categorical features



Attrition	No	Yes	Yes_pct
MaritalStatus			
Divorced	768	79	9.327037
Married	1541	225	12.740657
Single	940	322	25.515055

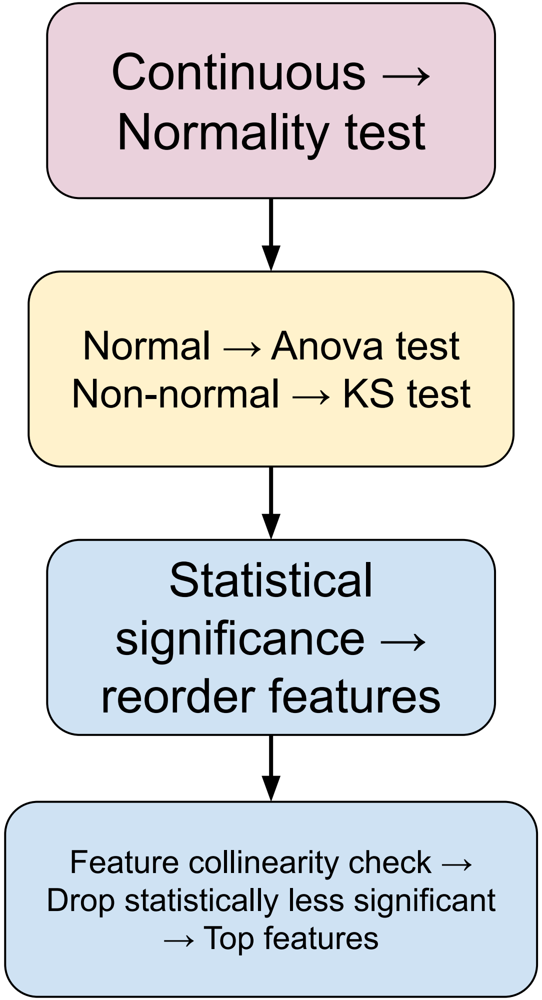
Attrition	No	Yes	Yes_pct
EnvironmentSatisfaction			
1	562	191	25.365206
2	646	115	15.111695
3	1035	163	13.606010
4	1006	157	13.499570

Attrition	No	Yes	Yes_pct
JobSatisfaction			
1	596	174	22.597403
2	619	124	16.689098
3	974	189	16.251075
4	1060	139	11.592994

Attrition	No	Yes	Yes_pct
WorkLifeBalance			
1	143	65	31.250000
2	767	157	16.991342
3	2005	330	14.132762
4	334	74	18.137255

Attrition	No	Yes	Yes_pct
EducationField			
Human Resources	43	27	38.571429
Life Sciences	1330	270	16.875000
Marketing	352	63	15.180723
Medical	1032	196	15.960912
Other	183	29	13.679245
Technical Degree	309	41	11.714286

Attrition	No	Yes	Yes_pct
JobRole			
Healthcare Representative	291	49	14.411765
Human Resources	121	19	13.571429
Laboratory Technician	570	110	16.176471
Manager	235	35	12.962963
Manufacturing Director	341	44	11.428571
Research Director	166	52	23.853211
Research Scientist	638	142	18.205128
Sales Executive	699	145	17.180095
Sales Representative	188	30	13.761468



Continuous → Normality test

Normal → Anova test
Non-normal → KS test

Statistical significance → reorder features

Feature collinearity check → Drop statistically less significant → Top features

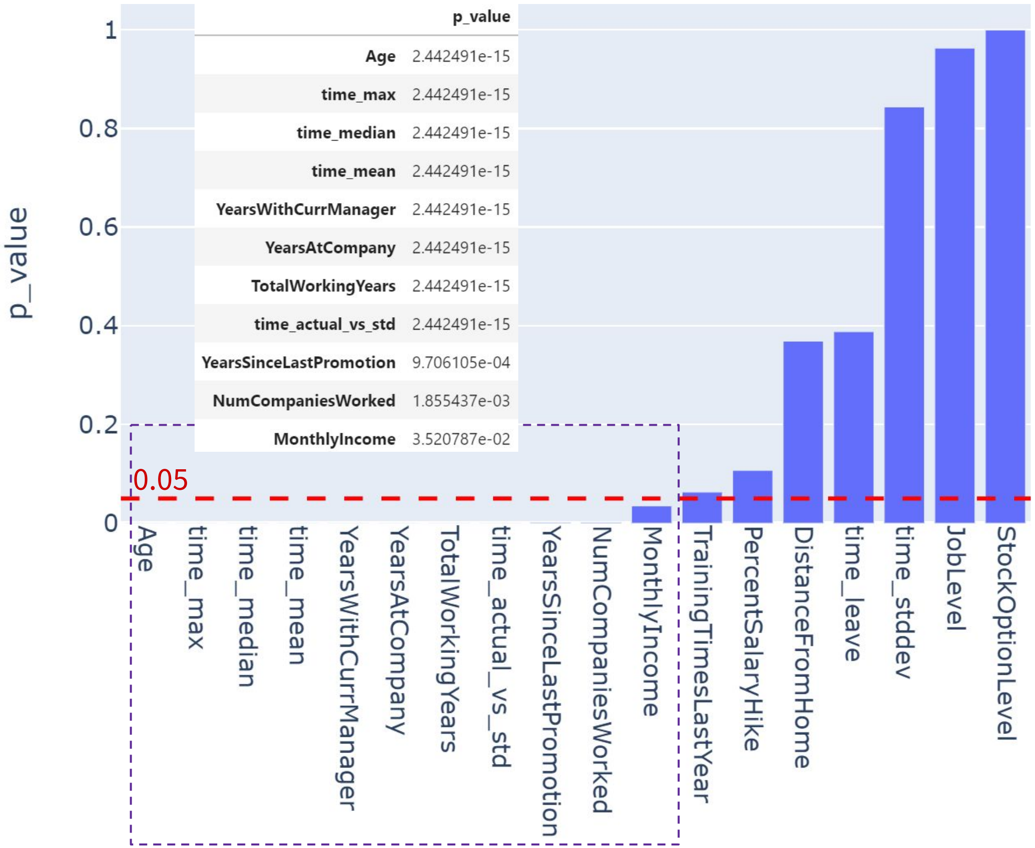
Statistical feature selection EDA *(1 of 2): Continuous*

Feature selection

1. Continuous features with statistically significant correlation to response - Age, Time: Max, Median, Mean, YearswithCurrmanager, YearsAtCompany, TotalWorkingYears, Time: Delta-to-Std, YearsSinceLastPromotion, NumCompaniesWorked & MonthlyIncome *(fig 8)*
2. Time features that correlated to each other, dropped based on statistical significance - Median, Mean & Delta-to-Std *(fig 9)*

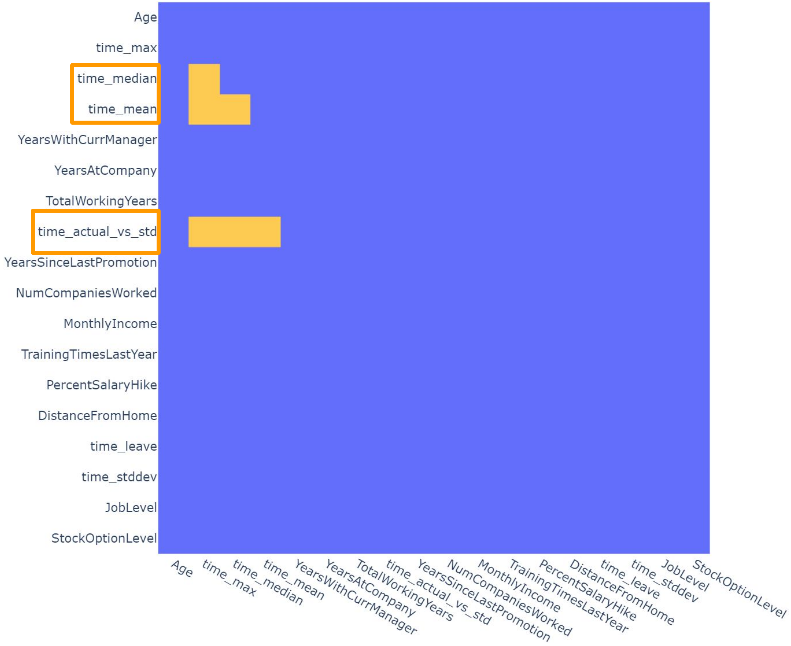
Statistical feature selection EDA *(2 of 2): Continuous*

fig 8. Continuous feature importance



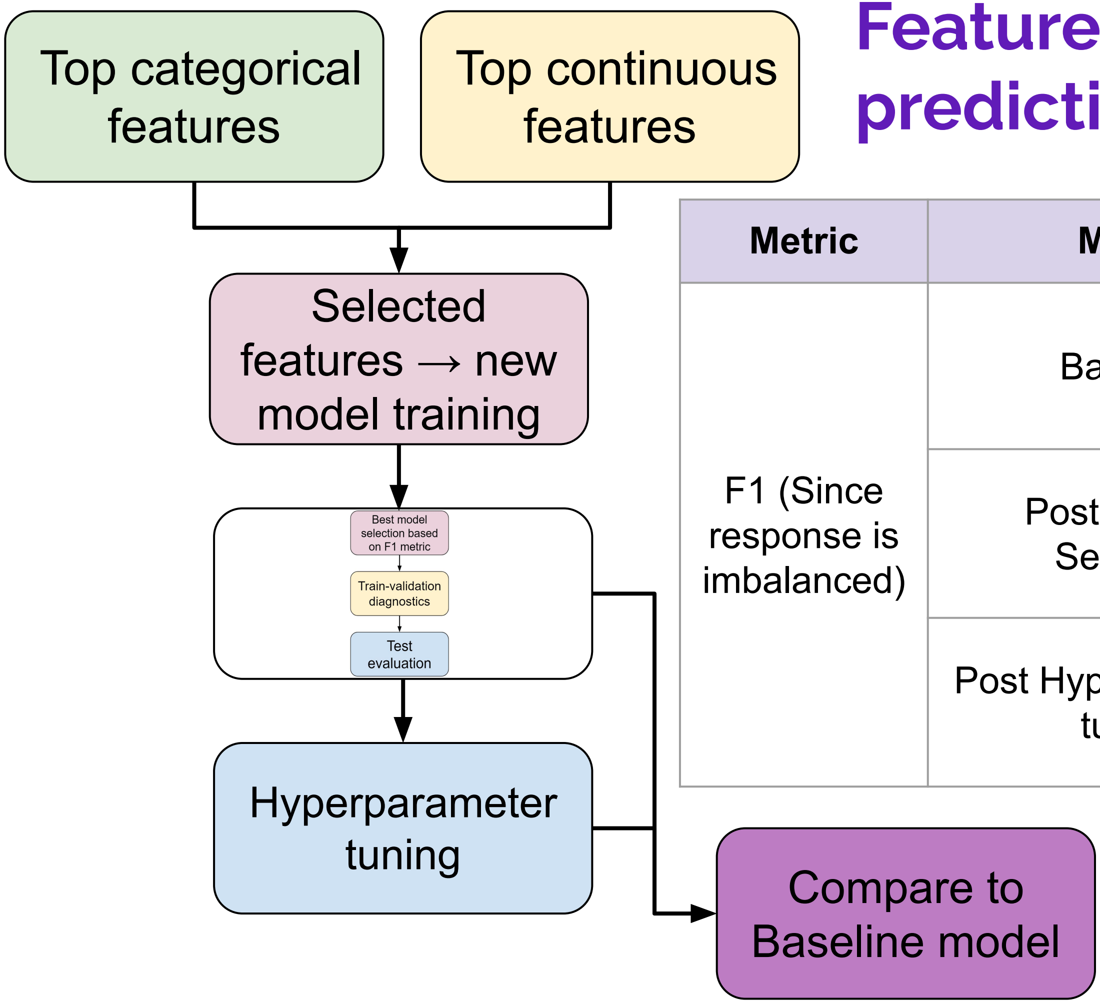
	p_value
Age	2.442491e-15
time_max	2.442491e-15
time_median	2.442491e-15
time_mean	2.442491e-15
YearsWithCurrManager	2.442491e-15
YearsAtCompany	2.442491e-15
TotalWorkingYears	2.442491e-15
time_actual_vs_std	2.442491e-15
YearsSinceLastPromotion	9.706105e-04
NumCompaniesWorked	1.855437e-03
MonthlyIncome	3.520787e-02

fig 9. Collinearity heatmap



Model
Post feature selection +
Hyperparameter tuning

Feature selected resignation prediction model *(1 of 4)*

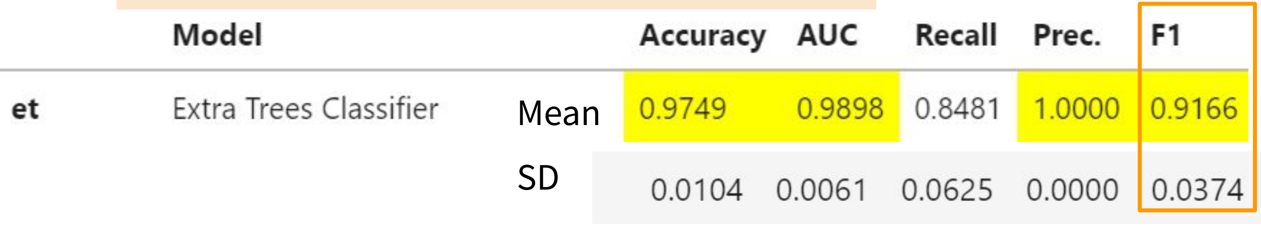

Top categorical features
Top continuous features
Selected features → new model training
Best model selection based on F1 metric
Train-validation diagnostics
Test evaluation
Hyperparameter tuning
Compare to Baseline model

Metric	Model	Split	Value	Fig
F1 (Since response is imbalanced)	Baseline	Train (10 folds)	0.911	5.1
		Test (holdout)	0.86	5.2
	Post Feature Selection	Train (10 folds)	0.917	10.1
		Test (holdout)	0.82	10.2
	Post Hyperparameter tuning	Train (10 folds)	0.913	11
		Test (holdout)	0.66	12

10 fold F1 SD: 0.03

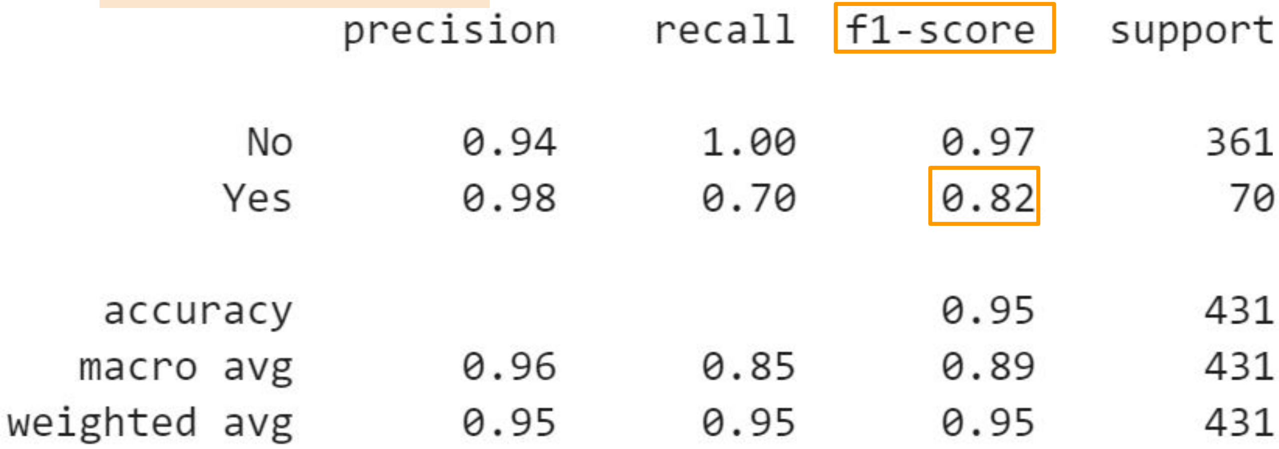
Feature selected resignation prediction model *(2 of 4)*

fig 10.1. Best model avg 10-fold CV metrics



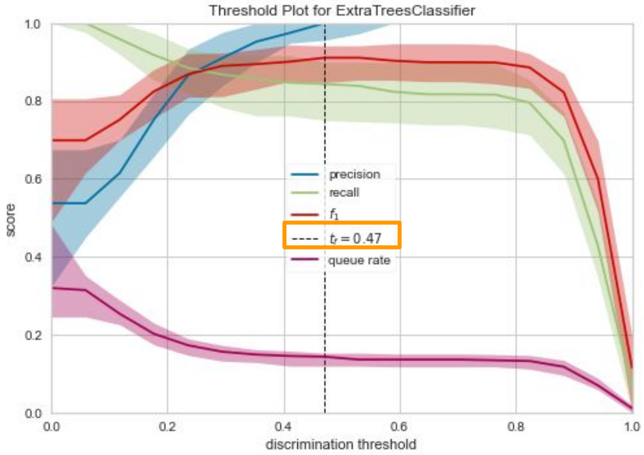
	Model		Accuracy	AUC	Recall	Prec.	F1
et	Extra Trees Classifier	Mean	0.9749	0.9898	0.8481	1.0000	0.9166
		SD	0.0104	0.0061	0.0625	0.0000	0.0374

fig 10.2. Test metrics



```
              precision    recall  f1-score   support

          No       0.94      1.00      0.97       361
         Yes       0.98      0.70      0.82        70

    accuracy                           0.95       431
   macro avg       0.96      0.85      0.89       431
weighted avg       0.95      0.95      0.95       431
```

fig 10.3 Model optimal threshold



Feature selected resignation prediction model *(3 of 4)*

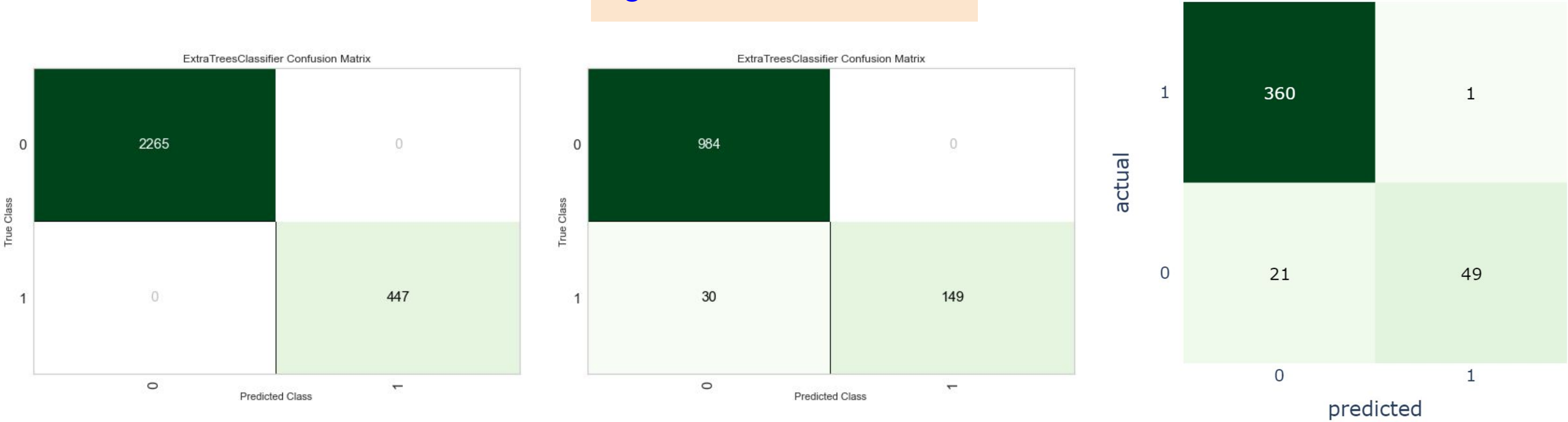


fig 10.4. Confusion matrix

True Class \ Predicted Class	0	1
0	2265	0
1	0	447

training

True Class \ Predicted Class	0	1
0	984	0
1	30	149

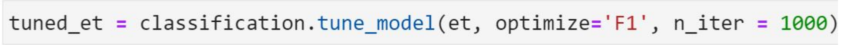
validation

actual \ predicted	0	1
1	360	1
0	21	49

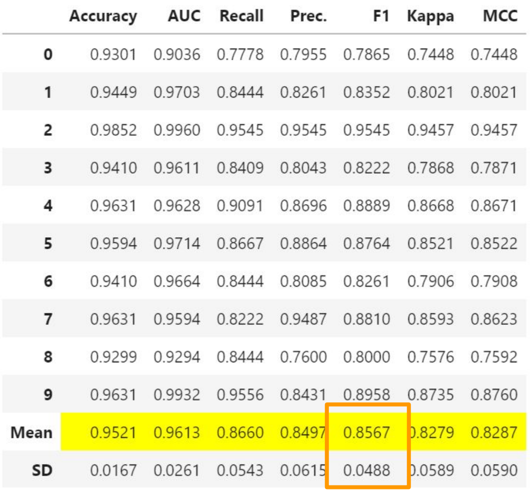
testing

Feature selected resignation prediction model *(4 of 4): Hyperparameter tuning*

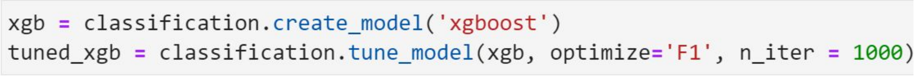
fig 11. Hyperparameter tuned model 10-fold CV avg metrics



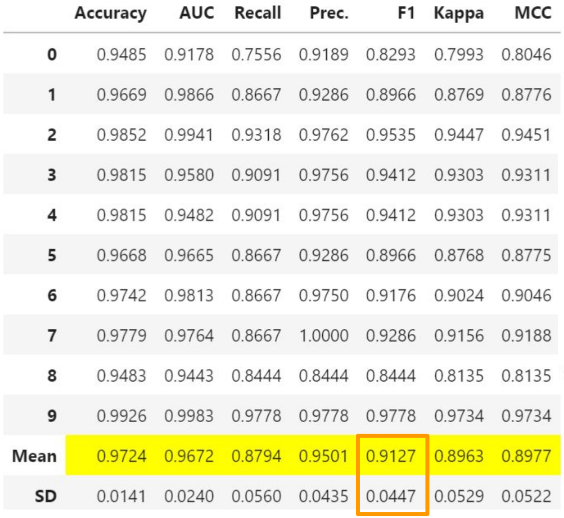
```
tuned_et = classification.tune_model(et, optimize='F1', n_iter = 1000)
```



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9301	0.9036	0.7778	0.7955	0.7865	0.7448	0.7448
1	0.9449	0.9703	0.8444	0.8261	0.8352	0.8021	0.8021
2	0.9852	0.9960	0.9545	0.9545	0.9545	0.9457	0.9457
3	0.9410	0.9611	0.8409	0.8043	0.8222	0.7868	0.7871
4	0.9631	0.9628	0.9091	0.8696	0.8889	0.8668	0.8671
5	0.9594	0.9714	0.8667	0.8864	0.8764	0.8521	0.8522
6	0.9410	0.9664	0.8444	0.8085	0.8261	0.7906	0.7908
7	0.9631	0.9594	0.8222	0.9487	0.8810	0.8593	0.8623
8	0.9299	0.9294	0.8444	0.7600	0.8000	0.7576	0.7592
9	0.9631	0.9932	0.9556	0.8431	0.8958	0.8735	0.8760
Mean	0.9521	0.9613	0.8660	0.8497	0.8567	0.8279	0.8287
SD	0.0167	0.0261	0.0543	0.0615	0.0488	0.0589	0.0590

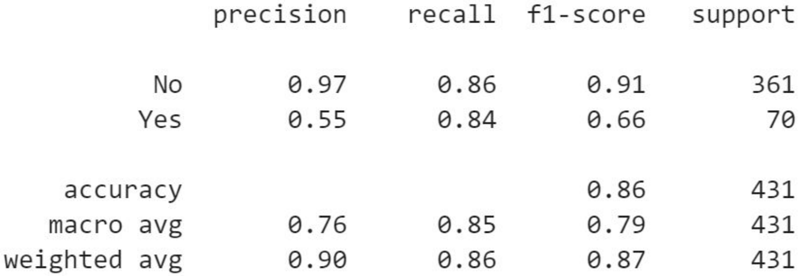


```
xgb = classification.create_model('xgboost')
tuned_xgb = classification.tune_model(xgb, optimize='F1', n_iter = 1000)
```



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9485	0.9178	0.7556	0.9189	0.8293	0.7993	0.8046
1	0.9669	0.9866	0.8667	0.9286	0.8966	0.8769	0.8776
2	0.9852	0.9941	0.9318	0.9762	0.9535	0.9447	0.9451
3	0.9815	0.9580	0.9091	0.9756	0.9412	0.9303	0.9311
4	0.9815	0.9482	0.9091	0.9756	0.9412	0.9303	0.9311
5	0.9668	0.9665	0.8667	0.9286	0.8966	0.8768	0.8775
6	0.9742	0.9813	0.8667	0.9750	0.9176	0.9024	0.9046
7	0.9779	0.9764	0.8667	1.0000	0.9286	0.9156	0.9188
8	0.9483	0.9443	0.8444	0.8444	0.8444	0.8135	0.8135
9	0.9926	0.9983	0.9778	0.9778	0.9778	0.9734	0.9734
Mean	0.9724	0.9672	0.8794	0.9501	0.9127	0.8963	0.8977
SD	0.0141	0.0240	0.0560	0.0435	0.0447	0.0529	0.0522

fig 12. Hyperparameter tuned model test metrics (xgb)



```
              precision    recall  f1-score   support

          No       0.97      0.86      0.91       361
         Yes       0.55      0.84      0.66        70

    accuracy                           0.86       431
   macro avg       0.76      0.85      0.79       431
weighted avg       0.90      0.86      0.87       431
```

Summary

Highlights

Strong baseline model

- 91% F1_score
- 100% Precision

- Highly predictive model constructed with feature engineering and auto-ML library Pycaret
- Further feature selection, hyperparameters tuning improved model marginally

Clarifications

- Missing feature
- Attrition ambiguity

- Missing - RelationshipSatisfaction
- No resignation dates
- Employee worktimes are recorded throughout 2015

Explorations

- Model based missing value imputation
- Exit interview comments
- Survey comments

- Exploring Iterative imputer to estimate feature with missing values based on other influencing features
- Remodeling with missed feature
- NLP: analyzing exit interview & survey comments to understand resignee archetypes

Thank You!