Please insert Client/ Partner Logo here

Identifying potentially matching user's data

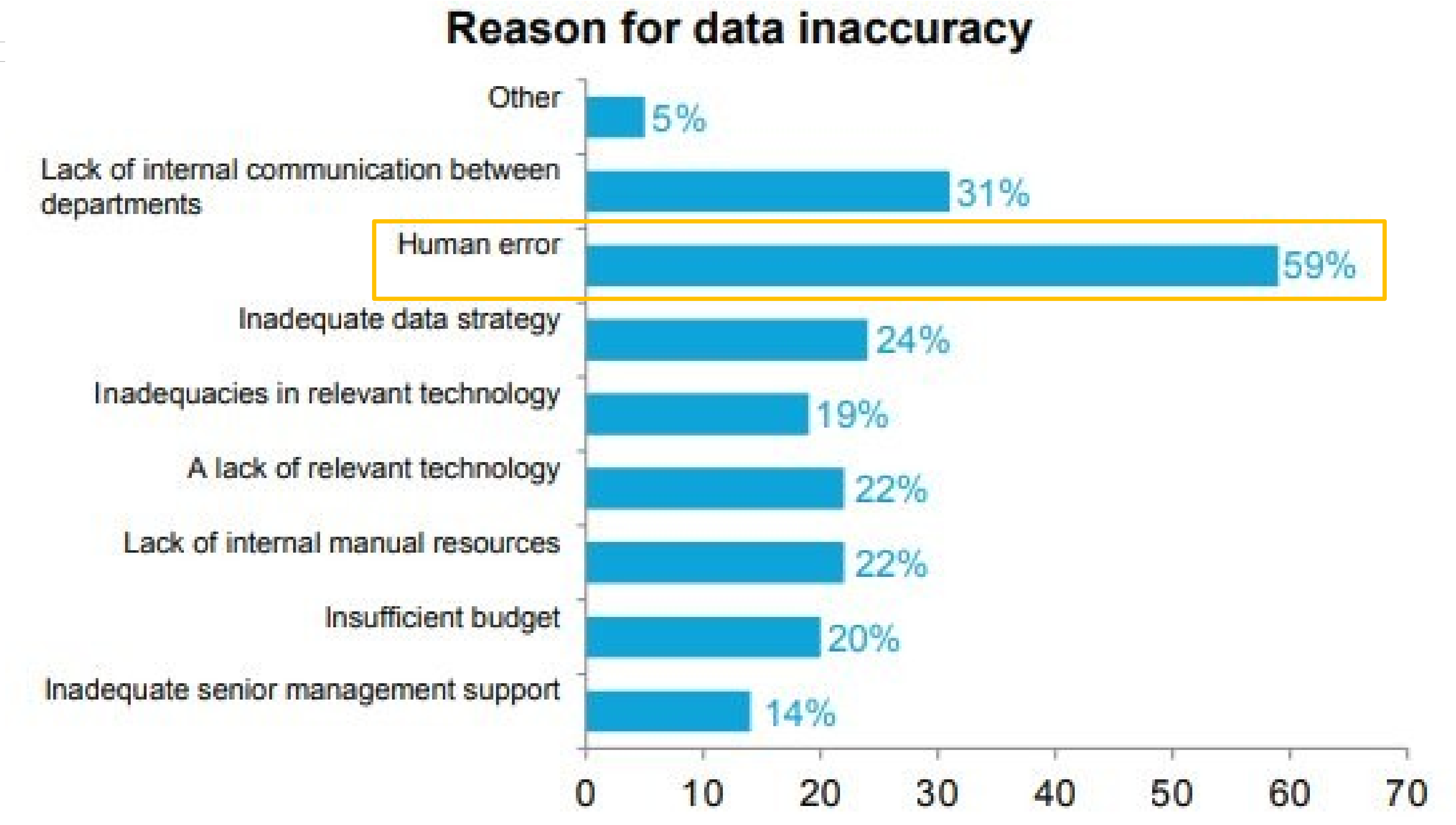
Proposed method using a machine learning model

WE INNOVATE TO SIMPLIFY INSURANCE

Risks/Issues with duplicate data

- Lack of a single customer view
- Costs and lost productivity
- Brand trust and credibility are put at risk
- Customer relationships and experience are impaired
- Valuable and expensive data storage space is sacrificed
- Duplicate records make it harder, if not impossible to comply with regulatory (GDPR etc.,)
- Leads to incorrect MI reports

Reasons for duplicate data



Duplicates may occur during data migrations as well when multiple legacy data sources are moved to a single data source

Challenges to identify duplicate user data

Difficult to know whether multiple data records are related to same person/entity unless all the essential details are identical

Types of data similarities in Names, Addresses and other user sensitive/identification data

Phonetic similarity
Textual similarity
Nickname
Missing spaces/hyphens
Initials
Name swap
Different name split
Truncated name
Missing name
Maiden name addition

Rule-based vs Machine Learning

Approach	Rule-based	Machine Learning
Operation	Rule generation, rule-based identification	Feature extraction, training and test
Model	Empirically derived rules	Automatic generated
Labelled data	No	Yes (for training)
Manual effort	Needed for generating rules	Minimal
Major cost	Rule tuning by human	Pairing operation
Flexibility	Low (cannot recognizes beyond rules)	High (can adopt to subtle cases)

Machine Learning Model Implementation

Aim is to build a supervised machine learning model using a classification algorithm/s and train it to infer a given source data record potentially matches with a target data record or not

Combination of the following features/ columns used as a data record

- First name
- Surname
- Gender
- DOB
- NINO
- Postcode

Python language library FuzzyWuzzy identified to check the text similarities of First name and Surname features/columns between two records. It has following methods which will compare text and give a score between 0 to 100

- Ratio
- PartialRatio
- TokenSortRatio
- TokenSetRatio
- WRatio

Machine Learning Model Implementation

32 (2^5) data scenarios/combinations used as a base to prepare training data set for the model to learn/train duplicate data vs non duplicate data. 6000+ rows of data samples prepared to cover the identified scenarios

The source (records to match) and target (record/s to match against) data records are maintained in a CSV file and fed to the model

The matching of Gender, DOB, NINO and Postcode between Source and Target records done using string comparison and the output will be either 1 (match) or 0 (no match). The string matching of First name and Surname done using Python FuzzyWuzzy library and the output will vary from 0 to 100

The matching results between all Source and Target data records and the actual values (i.e., duplicate or not) fed to the model in 80-20 ratio i.e., to train the model using 80% data and test its accuracy with the remaining 20% data

The following 5 different algorithms chosen to train the model parallelly which will help to choose a high accuracy model after trained and tested

Random Forest, AdaBoost, Decision Tree, KNeighbors & GaussianNB

Data scenarios

Name Match	Gender Match	DOB Match	NINO Match	Postcode Match	Duplicate (Label)
Y	Y	Y	Y	Y	Y
Y	Y	Y	Y	N	Y
Y	Y	N	Y	Y	Y
Y	N	Y	Y	Y	Y
N	Y	Y	Y	Y	Y
N	Y	Y	Y	N	Y
N	Y	N	Y	Y	Y
N	N	Y	Y	Y	Y
Y	Y	Y	N	Y	N
Y	Y	Y	N	N	N
Y	Y	N	Y	N	N
Y	Y	N	N	Y	N
Y	Y	N	N	N	N
Y	N	Y	Y	N	N
Y	N	Y	N	Y	N
Y	N	Y	N	N	N
Y	N	N	Y	Y	N
Y	N	N	Y	N	N
Y	N	N	N	Y	N
Y	N	N	N	N	N
N	Y	Y	N	Y	N
N	Y	Y	N	N	N
N	Y	N	Y	N	N
N	Y	N	N	Y	N
N	Y	N	N	N	N
N	N	Y	Y	N	N
N	N	Y	N	Y	N
N	N	Y	N	N	N
N	N	N	Y	Y	N
N	N	N	Y	N	N
N	N	N	N	Y	N
N	N	N	N	N	N

Sample data preparation

Duplicate (Label)	Surname1	Firstname1	Gender1	DOB1	NINO1	Postcode1	Surname2	Firstname2	Gender2	DOB2	NINO2	Postcode2
1	Stefan	Andrews	F	10/10/2010	FYC88TV2	K9J 7AF	Stefan	Andrews	F	10/10/2010	FYC88TV2	K9J 7AF
1	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ
1	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ
1	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI
1	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	ane	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zne	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zae	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zan	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zane	ebb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
0	Stefan	Andrews	F	10/10/2010	FYC88TV23X	K9J 7AF	Stefan	Andrews	F	10/10/2010	FYC88TV2	K9J 7AF
0	Kacper	Carter	M	02/06/1986	IWK78LX93X	J39 7QQ	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ
0	Jay	Khan	F	30/08/1947	YYH59NR83X	UK5C 5DZ	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ
0	Francis	Dawson	M	21/05/1991	FJF25AX13X	RS9R 1GI	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI
0	Zane	Webb	F	25/07/1978	KIG76UL93X	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER

Test data

Duplicate (Label)	Surname1	Firstname1	Gender1	DOB1	NINO1	Postcode1	Surname2	Firstname2	Gender2	DOB2	NINO2	Postcode2
1	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
1	Vijay	R	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
1	Vijay	Ragot	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	Vijay	Kumar	M	30/01/1990	YUW76JI5	T1R 9JA	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	Vijay	Ragothaman	M	20/04/1975	QSF43SN1	DV8 4IF	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
1	Viji	Rugothuman	M	13/12/1998	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	Kumara	Guru	F	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH

Model output

Predicted	Actual	Surname1	Firstname1	Gender1	DOB1	NINO1	Postcode1	Surname2	Firstname2	Gender2	DOB2	NINO2	Postcode2
1	1	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	1	Vijay	R	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	1	Vijay	Ragot	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Vijay	Kumar	M	30/01/1990	YUW76JI5	T1R 9JA	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Vijay	Ragothaman	M	20/04/1975	QSF43SN1	DV8 4IF	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
1	1	Viji	Rugothuman	M	13/12/1998	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
1	0	Kumara	Guru	F	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH

More data

Duplicate (Label)	Surname1	Firstname1	Gender1	DOB1	NINO1	Postcode1	Surname2	Firstname2	Gender2	DOB2	NINO2	Postcode2
1	Stefan	Andrews	F	10/10/2010	FYC88TV2	K9J 7AF	Stefan	Andrews	F	10/10/2010	FYC88TV2	K9J 7AF
1	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ
1	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ
1	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI
1	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	ane	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zne	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zae	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zan	Webb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
1	Zane	ebb	F	25/07/1978	MXKIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
0	Ayaz	Welch	N	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
0	Malia	Melton	N	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
0	Manraj	Wills	N	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
0	Karim	Rosas	N	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
0	Aliza	Jacobson	N	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER
0	Viktoria	Schofield	N	25/07/1978	KIG76UL9	E9 8ER	Zane	Webb	F	25/07/1978	KIG76UL9	E9 8ER

Model output

Predicted	Actual	Surname1	Firstname1	Gender1	DOB1	NINO1	Postcode1	Surname2	Firstname2	Gender2	DOB2	NINO2	Postcode2
1	1	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	1	Vijay	R	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	1	Vijay	Ragot	M	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Vijay	Kumar	M	30/01/1990	YUW76JI5	T1R 9JA	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Vijay	Ragothaman	M	20/04/1975	QSF43SN1	DV8 4IF	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
1	1	Viji	Rugothuman	M	13/12/1998	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Kumara	Guru	F	29/03/1985	ABCDEFG	GL7 1JH	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Kacper	Carter	M	02/06/1986	IWK78LX9	J39 7QQ	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Jay	Khan	F	30/08/1947	YYH59NR8	UK5C 5DZ	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH
0	0	Francis	Dawson	M	21/05/1991	FJF25AX1	RS9R 1GI	Vijay	Ragothaman	M	29/03/1985	ABCDEFG	GL7 1JH

Model Quality

		Predicted Class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive (TP)	False Positive (FP)
	Class = No	False Negative (FN)	True Negative (TN)

Accuracy

Calculated as (TP + TN) / (TP + TN + FP + FN)

Model achieved **93%** accuracy

Precision

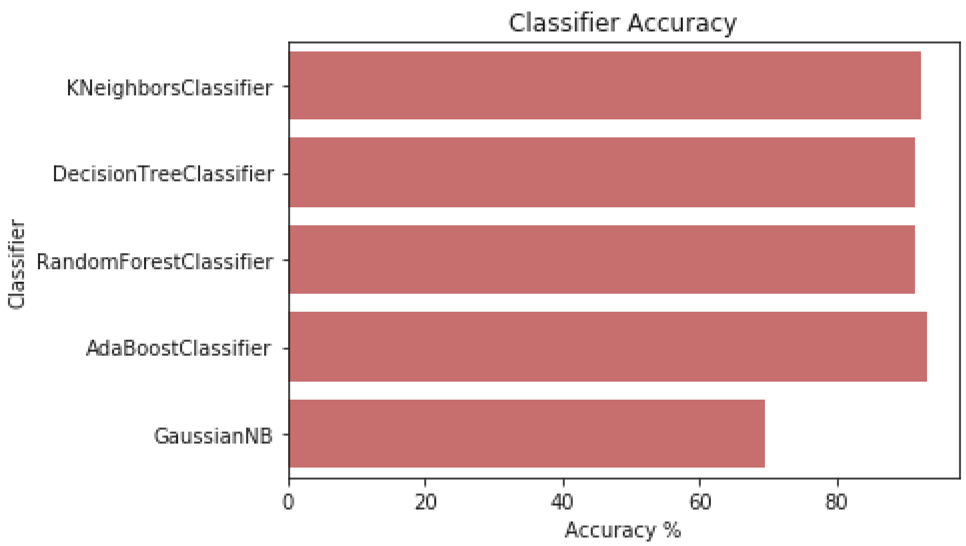
Calculated as TP / (TP + FP)

Model achieved **95%** precision

Important quality measure of a model because it will be worse to give positive match between two people who aren’t really the same person than missing a match between two people who are actually same person

Model Quality

Accuracy



Precision

	precision	recall	f1-score	support
0	1.00	0.91	0.95	978
1	0.76	1.00	0.87	285
avg / total	0.95	0.93	0.93	1263

Few Use cases

- Prompt user while logging Client details in UI
- Client records matching to identify a Golden record
- Anti Money Laundering screening for regulatory compliance
- Identifying and cleaning duplicates in existing DBs
- Sense checks during data migrations

Challenges / Learnings

Data Scenarios

Training and preparation

Feature extraction

Learning and testing iteration

intellect SEEC™

DISTRIBUTION SOLUTIONS

PIPELINE MANAGEMENT

NEEDS ANALYSIS

PROFILE MANAGEMENT

QUOTE & ILLUSTRATION

e-APP

CUSTOMER SERVICING

PRODUCT CONFIGURATOR

UNDERWRITING SOLUTIONS

UNDERWRITING WORKSTATION

RISK ANALYTICS

CLAIMS SOLUTIONS

INTELLECT CLAIMS

LIFE CLAIMS IN THE CLOUD

BUSINESS COMPONENT SOFTWARE

BUSINESS COMPONENTS

WE INNOVATE TO SIMPLIFY INSURANCE