Fraud Prediction using AutoAI

Preface

Overview

Automation and artificial intelligence (AI) are transforming businesses and will contribute to economic growth via contributions to productivity. They will also help address challenges in areas of healthcare, technology & other areas. At the same time, these technologies will transform the nature of work and the workplace itself. In this code pattern, we will focus on building state of the art systems for churning out predictions which can be used in different scenarios. We will try to predict fraudulent transactions which we know can reduce monetary loss and risk mitigation. The same approach can be used for predicting customer churn, demand and supply forecast and others. Building predictive models require time, effort and good knowledge of algorithms to create effective systems which can predict the outcome accurately. With that being said, IBM has introduced AutoAI which will automate all the tasks involved in building predictive models for different requirements. We will get to see how AutoAI can churn out great models quickly which will save time and effort and aid in faster decision-making process.

Industry Use-case

A. Fraud detection in the insurance business

Headquartered in Aurora, just outside of Chicago, Northeast Insurance Company, (NIC) employs over 5000 people across the continental United States.

During its fifty-year history, NIC has been struggling to detecting potentially fraudulent activity and has turned to IBM for their data science and AI offerings to predict fraud with insurance claims, before the claim is settled.

B. Business challenge story

The global Fraud Detection and Prevention (FDP) market size is expected to grow from USD 20 billion to 63.5 billion by 2023, according to various analyst reports (i.e. "Fraud Detection and Prevention Market by solution). Predictive analytics segment is projected to be the largest contributor to the FDP market during the forecast period.

Predictive analytics solutions help enterprises identify the possibilities of fraud incidents by analyzing the current data. The solutions are used to identify potential threats, payment frauds, frauds in insurance processes, and credit/debit card frauds. Organizations are trying to impart these solutions for predicting fraud or suspicious activity and their pattern to help drastically reduce losses due to frauds.

Fraud analytics solutions employ sophisticated analytics and predictive modeling to identify potential fraud in real time during data entry, rather than during a later batch run after a transaction is complete. It can be applied to claims and underwriting fraud. Regional Tier 2 and 3 insurers are more likely to adopt SaaS-based point solutions for fraud analytics use cases. Larger insurers are implementing these solutions via professional services providers and system integrators.

Drivers usually sign a six-month policy with an auto insurance policy. Each month, or all at once, the driver pays a fee, or **premium**, to the company. There are a few things that determine the cost of the policy: the type of car insured (particularly its safety record and how expensive it is to repair) the driver's record (the more speeding tickets the driver has incurred, the riskier he is) and even age (teenagers cost more to insure because they're less experienced drivers, and therefore a bigger risk.) Lower-cost premiums are enjoyed by drivers with fewer accidents and tickets on their records, part-time drivers, people who take driver education courses, and families with multiple cars.

PAIN POINTS

Information siloed, overload, difficult to see clearly
Using AI/ML for fraud detection is not new. However, typical organization contains multiple fraud departments, each with its own internal point-solution which monitors fraud for that specific channel, product, or fraud type. Structured and unstructured data collected internally and externally but very few of these point-solutions share data. Each uses varying analytical techniques across channels and transaction systems, which results in not having a complete view of risk exposures across the institution. Cannot see patterns or behaviors that would spark a concern that fraudulent activity is crossing multi-business lines because the observation space is too narrow.

Difficult to predict fraud
Rare occurrences create an imbalance in the classification of fraud detection models and makes detection challenging.

Shift to increased digital and mobile customer platforms led to transactions being executed more quickly, leaving banks and processors with less time to identify, counteract, and recover the underlying funds. As quickly as new technology is used to identify fraudsters, they themselves are identifying new ways of defrauding the bank. For instance, identity theft is mutating from card skimming to account takeovers (ATO). Synthetic identify, a scenario where fraudsters combine fragments of stolen or fake information to create a new identity and apply for financial products.

Cost of (near) real-time detection is high
Organizations need to identify anomalies accurately and efficiently at the level of accounts, merchants, cardholders and locations.

False positives require manual investigations through providing content analytics across primary internal and external data sources

Fraud detection – meaning detecting fraudulent behavior after it occurs – forcing companies to set aside money and resources for the inevitable losses they will incur, costing financial institutions millions of dollars and destroying the customer experience. Financial institutions need to get in front of the problem and focus on fraud prevention.

72%

cite fraud as a growing concern over the past 12 months and nearly **63%** report the same or higher levels of fraudulent losses over that same period

$44B

Worldwide losses due to fraud by 2025

25%

of declined sales transactions for e-commerce merchants were false positives.

Tools

- IBM Watson Studio: Analyze data using RStudio, Jupyter, and Python in a configured, collaborative environment that includes IBM value-adds, such as managed Spark.
- IBM Auto AI: The AutoAI graphical tool in Watson Studio automatically analyzes your data and generates candidate model pipelines customized for your predictive modeling problem.
- IBM Cloud Object Storage: An IBM Cloud service that provides an unstructured cloud data store to build and deliver cost effective apps and services with high reliability and fast speed to market. This code pattern uses Cloud Object Storage.

Understanding the Data

Let's begin by accessing the CSV file required for the following steps.

1. Download the CSV file: https://github.com/apischdo/msce-632/blob/main/AutoInsClaims.csv.

   It is recommended that you use Firefox, IE or Mozilla (Edge and Safari are not recommended). Once you access the Github page, click **Raw**. Then, from the ensuing page, right-click and save the file as CSV (change file type to **All file types**).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	HOUSEHOLD	DRIVER_ID	POLICY_ID	CLAIM_ID	INCIDENT_C/	DESCRIPTION	CLAIM_STAT	ODOMETER_	LOSS_EVENT	CLAIM_INIT_	POLICE_REP(	CLAIMS_AT_	LOSS_LOCAT	LOSS_LOCAT	CLAIM_AMO	FLAG_FOR_F	PRIMARY_DI	START_DATE	EXPIRY_DATI	MODEL_YEA	MAKE	M(
2	CH42335	XZJ2837	NW5567882	A-2017-UU9	3		1	157654.9	4/25/17	4/28/17	1	1	41.9021031	-87.755624	35765	1	XZJ2837	9/19/16	9/19/17	2010	Nissan	Se
3	IH49805	VVR6423	UR4864804	A-2018-FI48	3		3	226154.5	8/26/18	8/31/18	0	1	41.9635619	-87.731397	1909	0	VVR6423	6/14/17	6/14/18	2008	Dodge	Ra
4	AF28736	UQM2512	RR8595908	A-2016-ZG6	1		1	83968.6	1/7/16	1/11/16	0	2	41.7366016	-87.604968	25730	1	UQM2512	8/24/15	8/23/16	2009	Chevrolet	Eq
5	EF53594	YDT5591	RN5640634	A-2016-NG7	1		1	309570.3	12/11/16	12/18/16	0	1	41.9099253	-87.731557	40880	1	YDT5591	7/25/15	7/24/16	2002	Chevrolet	Eq
6	LD32277	ONM5465	YY1229530	A-2017-ZO8	3		3	136633.9	6/6/17	6/8/17	0	1	41.9237502	-87.789881	2130	0	JZZ5688	1/8/16	1/7/17	2010	Honda	CR
7	DM94074	GBU7751	XP3473763	A-2018-XB4	1		3	326514.1	3/6/18	3/19/18	0	1	41.9092571	-87.785057	1970	0	BUK6977	1/15/17	1/15/18	2004	Ford	F-
8	MD38210	CBR4335	US5444269	A-2017-XP75	2		3	58477.9	8/13/17	8/18/17	0	1	41.9283951	-87.796468	2290	0	CBR4335	7/8/16	7/8/17	2012	Chevrolet	Im
9	GL77908	HZF3884	XR1994270	A-2017-QY9	1		3	176476.9	2/10/17	2/18/17	0	1	41.8857155	-87.728376	2990	0	QAX1481	8/11/16	8/11/17	2010	Kia	Sp
10	BA26199	CSE9523	VP6368585	A-2018-LB81	3		3	277812.7	2/16/18	2/24/18	0	1	41.895054	-87.745662	1170	0	CSE9523	10/13/17	10/13/18	2006	Nissan	Se
11	EA38976	HFX7408	YP9758006	A-2016-QD3	5		3	190541.3	12/28/16	1/11/17	0	1	41.8919756	-87.61458	2090	0	HFX7408	6/23/15	6/22/16	2008	Kia	Sp
12	GB64343	PXE3728	XY6800348	A-2017-CK71	1		3	290975.1	8/12/17	8/25/17	0	1	41.9112234	-87.638656	2130	0	PXE3728	4/14/16	4/14/17	2005	Kia	Op
13	JG99629	OKH5337	ZK6994471	A-2018-WF1	3		1	159873.2	7/30/18	7/30/18	1	5	41.7358355	-87.667866	33040	1	OKH5337	2/22/17	2/22/18	2010	Hyundai	So
14	FH81231	RQZ1566	ZR4462879	A-2018-VT47	1		3	391866.9	2/22/18	2/26/18	0	1	41.7341643	-87.551211	2400	0	FSC3949	1/10/17	1/10/18	1998	Dodge	Du
15	KI98597	VRX5780	ZW4263453	A-2017-SA27	3		1	384000.6	2/6/17	2/6/17	0	3	41.7918521	-87.801378	21330	1	VRX5780	6/4/16	6/4/17	2000	Toyota	Co
16	CG61685	TIA1702	UW8176531	A-2018-EJ36	2		3	251644.8	7/10/18	7/18/18	1	1	42.0059308	-87.68066	3427	0	TIA1702	4/10/17	4/10/18	2007	Nissan	Ro
17	AB12181	OBD8151	RW8652538	A-2017-CT58	2		1	300279.3	7/29/17	8/10/17	0	1	41.6934005	-87.612322	35385.5	1	OBD8151	11/25/16	11/25/17	2004	Chevrolet	Ca
18	BJ63668	WDC2749	PQ3189850	A-2017-CU8	2		3	73419.5	12/30/17	1/3/18	0	1	41.8401446	-87.661339	1570	0	OZP7842	4/6/16	4/6/17	2010	Honda	HR
19	MG83193	XJB5149	XN9015815	A-2018-LF97	2		3	118923.9	1/16/18	1/28/18	0	1	41.9751749	-87.768344	2370	0	XJB5149	11/21/17	11/21/18	2007	Jeep	Re
20	IJ01709	CRP8660	SV1659058	A-2017-ML6	1		1	182681.3	4/25/17	5/7/17	1	1	41.8901726	-87.649992	26250	1	CRP8660	7/23/16	7/23/17	2009	Chevrolet	Eq
21	BC94182	PEQ6155	UQ8163328	A-2018-WZ4	5		1	160429.9	2/10/18	2/19/18	1	3	41.8406579	-87.725899	25830	1	PEQ6155	7/11/17	7/11/18	2011	Nissan	Ve
22	MF96553	PPI8400	OT0287150	A-2018-DI45	2		1	165363.7	1/23/18	1/24/18	1	4	41.8847051	-87.667046	23483	1	PPI8400	4/23/17	4/23/18	2010	Hyundai	Tu
23	BC90853	JOA9170	TW2172802	A-2016-RW6	5		3	240639.2	12/1/16	12/10/16	1	1	41.7361826	-87.64366	2173.5	0	JOA9170	8/20/15	8/19/16	2005	Chevrolet	Ex
24	AK42388	KNQ4268	PP7360289	A-2017-KO8	4		3	388591.3	7/12/17	7/12/17	1	1	41.7620304	-87.611696	1750	0	KNQ4268	4/17/16	4/17/17	2001	Dodge	Ra
25	HB16909	SRU2950	OY6516260	A-2016-CA74	1		1	59498.7	1/6/16	1/8/16	1	1	42.0064221	-87.67812	20310	1	SRU2950	4/18/15	4/17/16	2013	Kia	Op
26	FA97534	EHL9301	OS6708536	A-2018-PV57	1		3	94043.6	10/13/18	10/25/18	0	1	41.8136018	-87.704338	2810	0	BIN1474	10/1/17	10/1/18	2014	Chevrolet	Co
27	CF57572	CJY6373	QV4842191	A-2017-JI571	4		3	69305.4	7/19/17	7/19/17	1	1	41.9969237	-87.806945	2370	0	WKB8440	3/6/16	3/6/17	2014	Mazda	CX
28	ME37855	NCL2868	VT9807337	A-2018-CK21	5		3	207764.4	11/4/18	11/13/18	0	1	41.9522275	-87.727425	1320	0	NCL2868	1/14/17	1/14/18	2008	Chevrolet	Sil
29	LA79549	YWA1319	PU5439335	A-2017-FJ21	5		1	302710	5/31/17	6/12/17	0	5	41.9768951	-87.69228	30840	1	YWA1319	4/4/16	4/4/17	2001	Acura	MI
30	II25056	CVI8835	OJ5466096	A-2017-YI13	4		3	109080	3/1/17	3/13/17	0	3	41.8807808	-87.730232	1230.5	0	CVI8835	5/20/16	5/20/17	2012	Dodge	Ra
31	CK11297	HIR9014	RR5447110	A-2018-QW1	4		3	69640.2	9/5/18	9/11/18	0	1	41.8927529	-87.760709	1020	0	HIR9014	7/11/17	7/11/18	2015	Toyota	Pri
32	KJ99727	MDE6706	RQ6230412	A-2018-AC52	2		2	281332.9	12/14/18	12/19/18	1	1	41.9383457	-87.712307	2470	0	MDE6706	5/17/17	5/17/18	2004	Honda	CR
33	GF42714	NKZ8115	ST5381628	A-2016-RA2	1		1	176864.1	7/14/16	7/25/16	0	2	41.9064302	-87.648306	31020	1	NKZ8115	8/23/15	8/22/16	2008	Hyundai	Tu
34	BD67160	SAD8006	YS2739165	A-2018-CE38	3		3	223253.6	11/26/18	11/30/18	0	1	41.890782	-87.631252	1890	0	SAD8006	2/8/17	2/8/18	2009	Toyota	Co
35	MJ57549	OGG2359	TO5690517	A-2018-CX48	3		3	139853.9	7/20/18	7/20/18	0	1	41.8924442	-87.6372	1932	0	OGG2359	8/3/17	8/3/18	2012	Hyundai	Tu
36	FI15095	MWA3612	NO2626234	A-2016-VA7	1		3	210844.5	9/24/16	9/26/16	0	1	41.8181867	-87.743319	1730	0	MWA3612	7/25/15	7/24/16	2005	Kia	Fo
37	ID61421	EXX9506	YS0850770	A-2016-EL66	2		3	282275.4	8/6/16	8/20/16	0	3	41.8824105	-87.637211	2170	0	EXX9506	9/15/15	9/14/16	2004	Ford	E-
38	MA52163	LSI8871	QR4862688	A-2017-TP82	1		1	224281.5	12/27/17	12/29/17	1	2	41.931362	-87.723151	20380	1	DLO2085	6/27/16	6/27/17	2007	Dodge	Ra

AutoInsClaims

Open the file with Excel and notice that the columns are in string format and some clearly need to be numeric and that is just at a first glance. Let's take a moment and understand the significance or more aptly, the predictability of each of the columns. Clearly, you can tell some of that info will not help you in predicting fraud. We need to remove the noise from the signal.

Take a moment, and note the column names: which may prove relevant? which are simply not needed? What if you used the date columns to calculate lapsed days from time of accident until reporting it? Or noted if the claims are being filled too close to the policy expiration date? Was there a police report? Let’s look at expired licenses at the time of submitting the claim. What about clients with low mileage discounts (7500 per year) that do not have low mileage?

Take a moment and consider the table below bearing in mind the questions asked above, let’s discuss that in a group if feasible, because before long, you will be immersed in feature engineering activities.

As you may have noted, there is a column named: **FLAG_FOR_FRAUD_INV**. Think of this column as training data for a supervised machine learning system. The rows with a value of 1 (True) have been verified and classified as fraud. That is known, not a prediction. We are going to build a model using this “training set” for data to predict future behavioral patterns that may then be flagged as potential fraud on a new CSV file never seen before by Watson Studio.

Feature name	**R for remove** **P for predictable** **and why do you think that?**
HOUSEHOLD_ID	R
DRIVER_ID	R
POLICY_ID	R
CLAIM_ID	R
INCIDENT_CAUSE 1 = driver error 2 = natural causes 3 = other driver error 4 = crime 5 = other causes	K
DESCRIPTION	R
CLAIM_STATUS 1 = open 2 = approved 3 = paid 4 = flagged for fraud 5 = denied 6 = appeal	R
ODOMETER_AT_LOSS	K
LOSS_EVENT_TIME	K
CLAIM_INIT_TIME	K
POLICE_REPORT 1 = there was police report 0 = no police report	K
CLAIMS_AT_LOSS_DATE (# of claims per individual)	K
LOSS_LOCATION_LAT	K
LOSS_LOCATION_LONG	K
CLAIM_AMOUNT	K
FLAG_FOR_FRAUD_INV	K (THIS IS YOUR X-AXIS)
PRIMARY_DRIVER_ID	R
START_DATE	K
EXPIRY_DATE	K
MODEL_YEAR	R

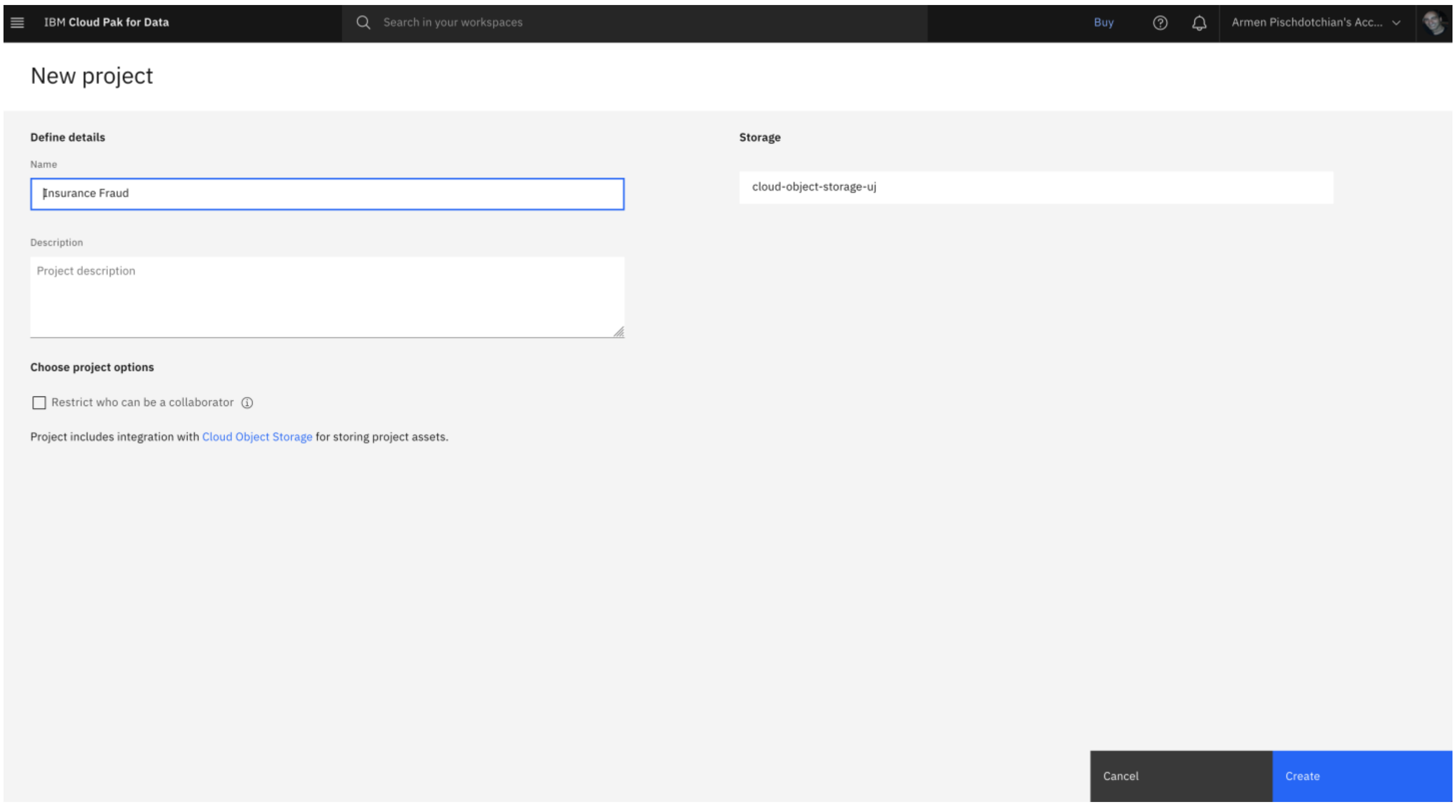
MAKE	R
MODEL	R
PLATE	R
COLOR	YOUR CHOICE
INITIAL_ODOMETER	K
LOW_MILEAGE_USE	K
FIRST_NAME	R
LAST_NAME	R
GENDER	K
BIRTHDATE	K
SSN	R
DRIVERS_LICENSE_ID	R
DRIVERS_LICENSE_EXPIRY	K
DRIVERS_LICENSE_STATE	WHAT DO YOU THINK?
DATE_AT_CURRENT_ADDRESS	K
CONTACT_NUMBER	R
EMAIL	R
COMMUTE_DISCOUNT	K

You are now ready for the detailed steps.

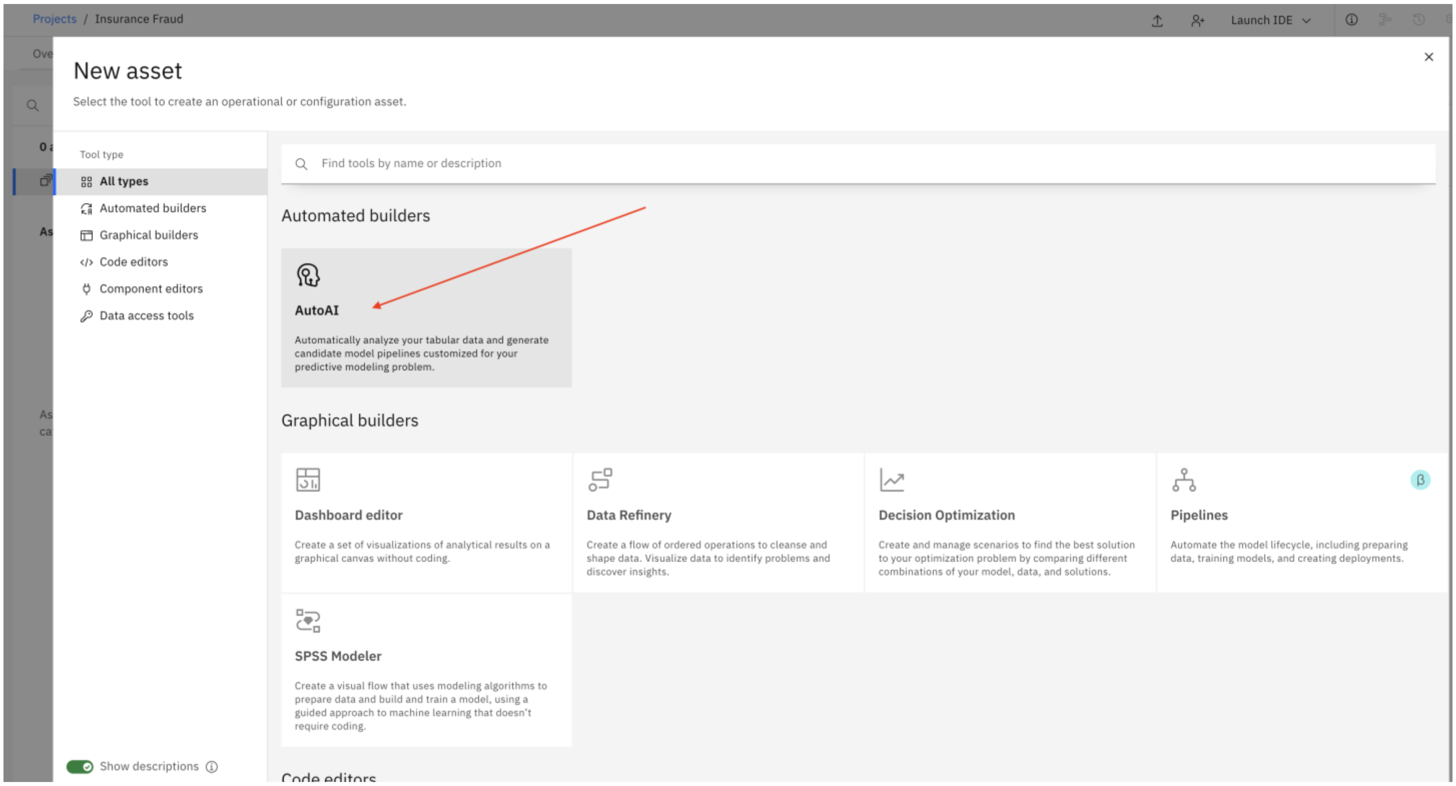
Build an AI model using AutoAI

Create a new Watson Studio project

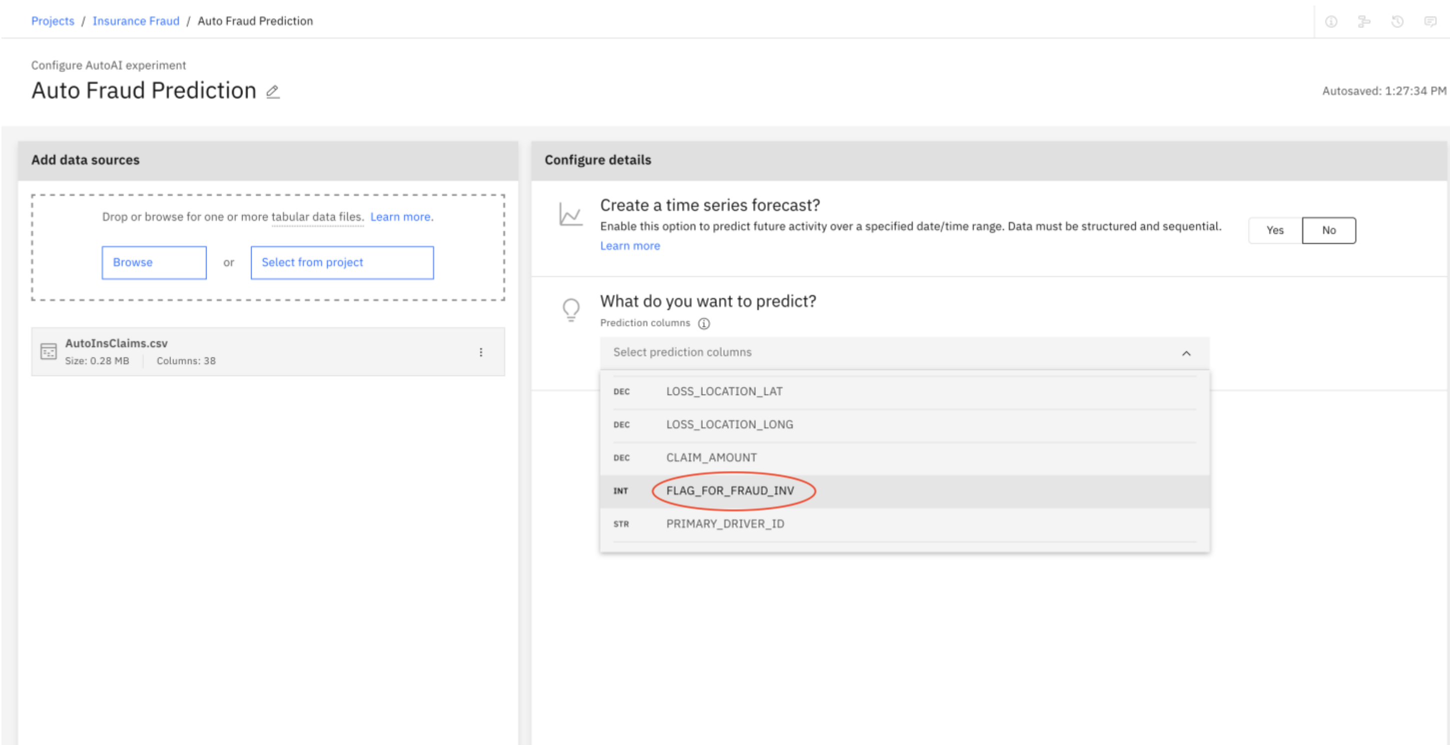
2. Provision Watson Studio.
3. Click **Launch in IBM Cloud Pak for Data**.
4. In the ensuing pop-up dialog, click **New Project** and click **Next**.
5. Define the project by giving a Name. Since you have already created the Object Storage, then it should appear under the Storage heading.



6. Click the **Assets** tab.
7. Click **New Assets** (blue box).

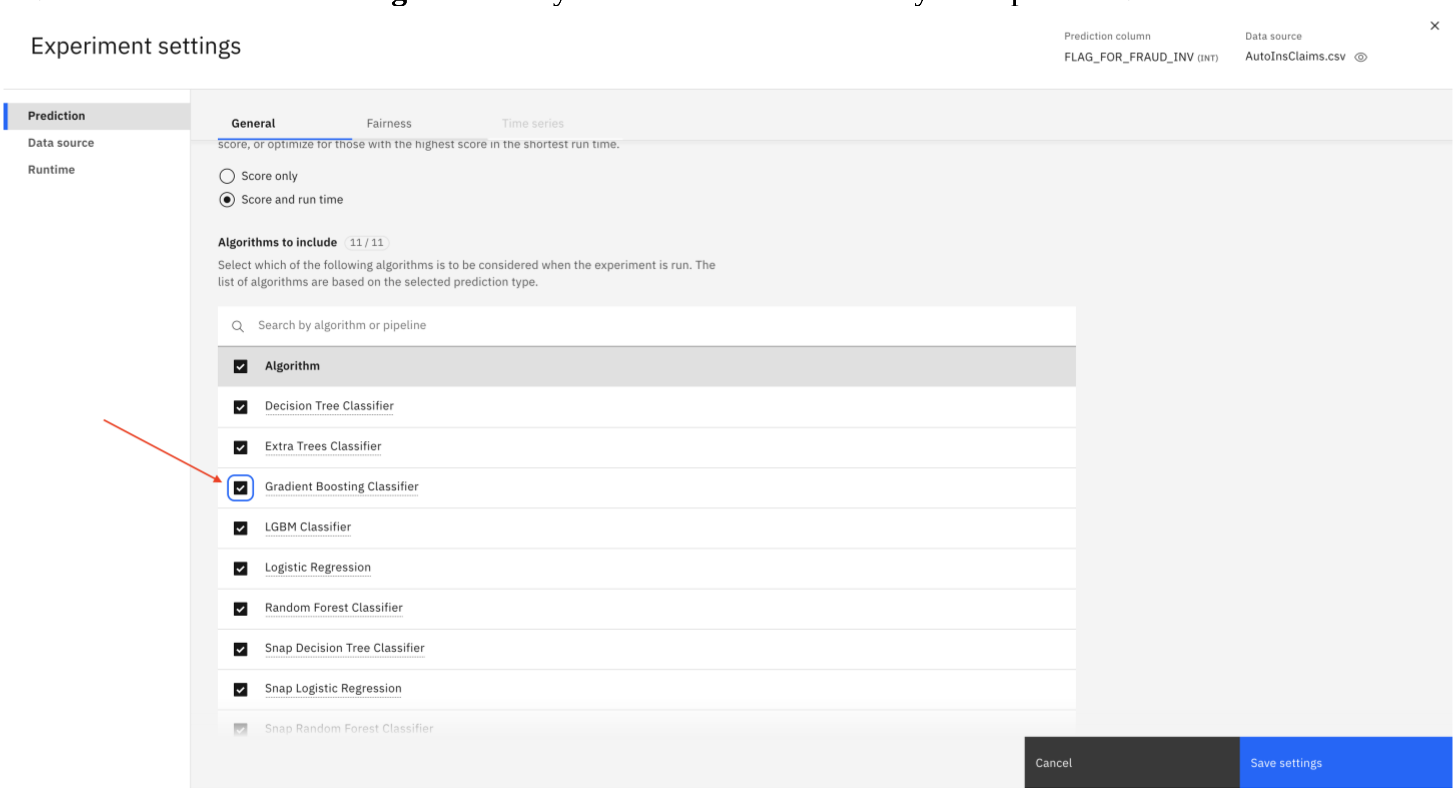


8. Select the **AutoAI** tile.
9. Specify a name for the experiment.
10. Click **Associate a Machine Learning service instance** to this project and select the Machine Learning service instance and click reload. If you do not have Machine Learning service instance, then follow the steps on your screen to provision the service.
11. Click **Create**.
12. Use the **Browse** button to upload your CSV file from your local drive.
13. Select **No**, since this is not a time series event.
14. Select **FLAG_FOR_FRAUD_INV** as the predictable column.
    a. Try out other features (column headings) and observe how the system recognizes Regression and Multiclassification as other potential approaches.
    b. Before you click **Run Experiment**, discuss with your team and instructor the nuances of what lurks in the Experiment Settings tab.

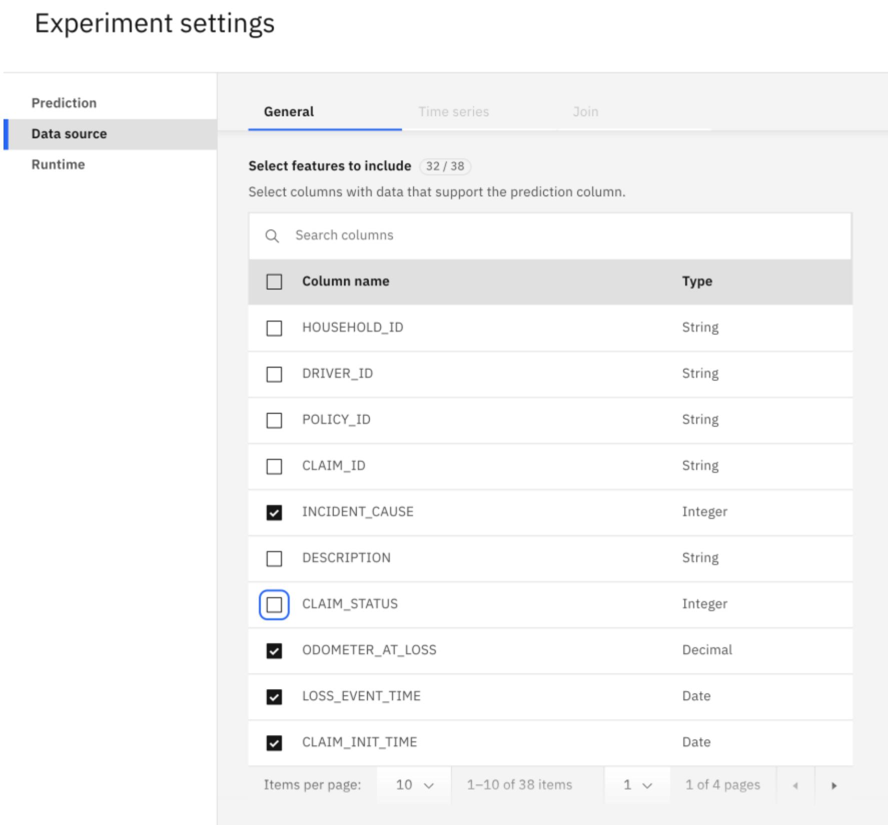


At this point you need to perform certain experiment settings.

15. Before you run the experiment, click **Experiment settings**.
16. Include **Gradient Boosting Classifier** yet another estimator to run your experiment.



17. Click the **Data source** tab and start unchecking the features that you deemed unpredictable from the table above (bear in mind there are 4 pages in this selection)

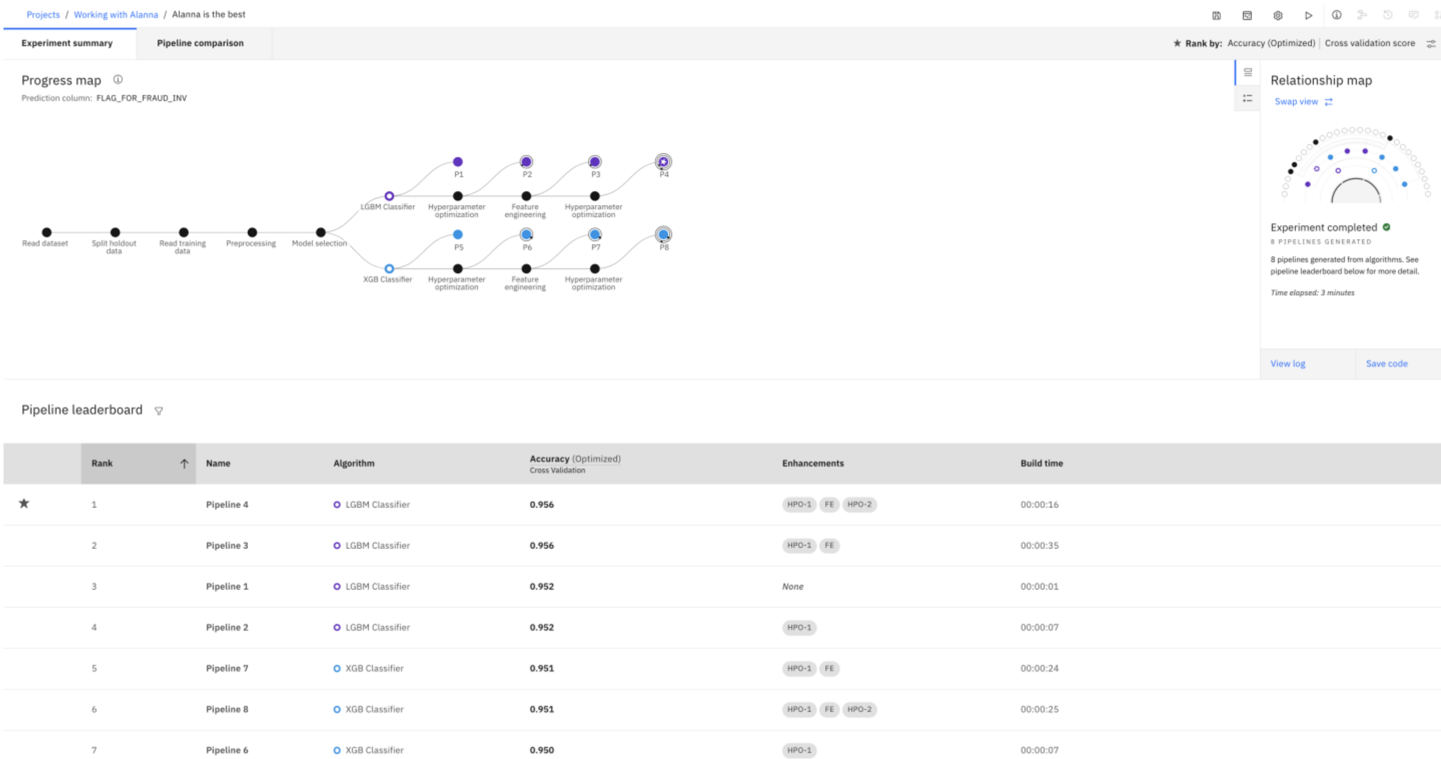


18. Save the settings
19. Run the Experiment.

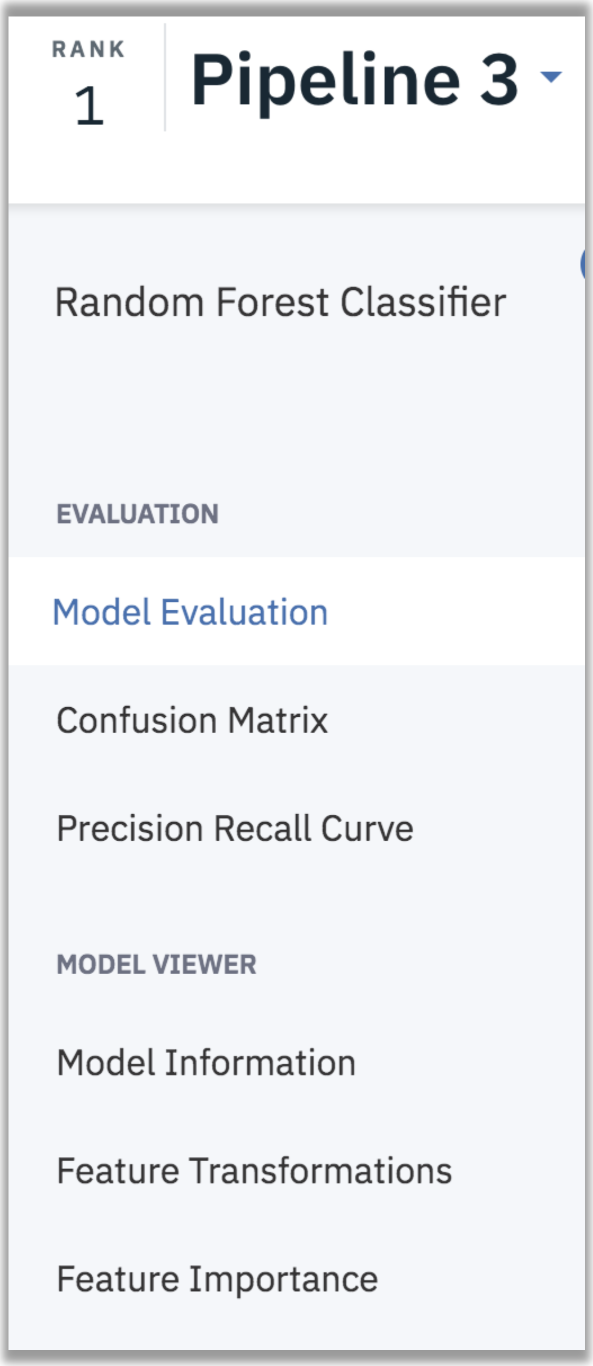
Analyze results

In this example, AutoAI experiment generated two pipelines. The duration of experiment depends completely on the size of the dataset. AutoAI selects the appropriate machine learning algorithm (in the fifth stage of the process under Model Selection) which is best suited for the dataset.

Each pipeline is run with different parameters, pipeline 2 is run on a sequence of HPO (hyper parameters optimization) & FE (feature engineering) whereas pipeline 4 includes HPO (hyper parameters optimization), FE (feature engineering) and a combination of both. All these are done on the fly! Isn't it amazing that we just have to sit and watch while AutoAI takes care of things for us and generates awesome machine learning models!! There's very minimal intervention required to get things going and in no time, we have the generated pipelines to choose from.



20. Click the highest-ranking pipeline (with the star) to see the evaluation metrics on the left side.



21. Click on model evaluation to review the performance of the model on the hold out sample and cross validation score.

We can observe that our model has done very well by scoring > 95% on Recall, average Precision scores & Area under the curve scores. These scores also mean that our model is able to remember and identify fraudulent transactions with great precision.

Model Evaluation Measures

	Holdout Score	Cross Validation Score
Accuracy	0.940	0.919
Area Under ROC Curve	0.982	0.970
Precision	0.939	0.978
Recall	0.958	0.881
F_1 Measure	0.948	0.926
Average Precision	0.989	0.979

These values are merely an example, and your values will be different. This is a probabilistic system.