أكاديمية سدايا
SDAIA Academy

Project report : Classification model
CAR INSURANCE FRAUD DETECTIONS

By

Hussain Alhadab

Ahmed Alonaizi

Mohammed Alhamoud

Feras Alyahya

Data Science Bootcamp SDAIA Academy

• Problem:

With the large number of traffic accidents insurance companies receive claims for financial compensation to the beneficiaries, and with the many claims appear multiple fraud cases and undeserved financial claims, so insurance companies face difficulty in identifying and detecting fraud in car accidents, insurance companies need a solution to help them detect fraud and identify the factors and causes of fraud, based on all factors related to the accident .

• Solution :

So in this project, we are going to develop a classification model to detect fraud.

• Data Description:

This Dataset contains **34 columns** and more than **11000 record**

The following table will explain the dataset in detail :

Columns	Type
Month	String
WeekOfMonth	Integer
DayOfWeek	String
Make	String
AccidentArea	String
DayOfWeekClaimed	String
MonthClaimed	String
WeekOfMonthClaimed	Integer
Sex	String
MaritalStatus	String
Age	Integer
Fault	String
PolicyType	String
VehicleCategory	String

VehiclePrice	**Integer**
FraudFound_P	Integer
PolicyNumber	Integer
RepNumber	Integer
Deductible	Integer
DriverRating	Integer
Days_Policy_Accident	Integer
Days_Policy_Claim	Integer
PastNumberOfClaims	Integer
AgeOfVehicle	Integer
AgeOfPolicyHolder	Integer
PoliceReportFiled	String
WitnessPresent	String
AgentType	String
NumberOfSuppliments	Integer
AddressChange_Claim	Integer
NumberOfCars	Integer
Year	Integer
BasePolicy	String
ClaimSize	Integer

• Tools:

Tools	Description
Jupyter notebook	Contains cells of Python code and human-readable text
pandas	The library is written in Python for data manipulation and analysis
sklearn	Software machine learning library for the Python programming language
Matplotlib	Matplotlib is a plotting library for Python

• Results:

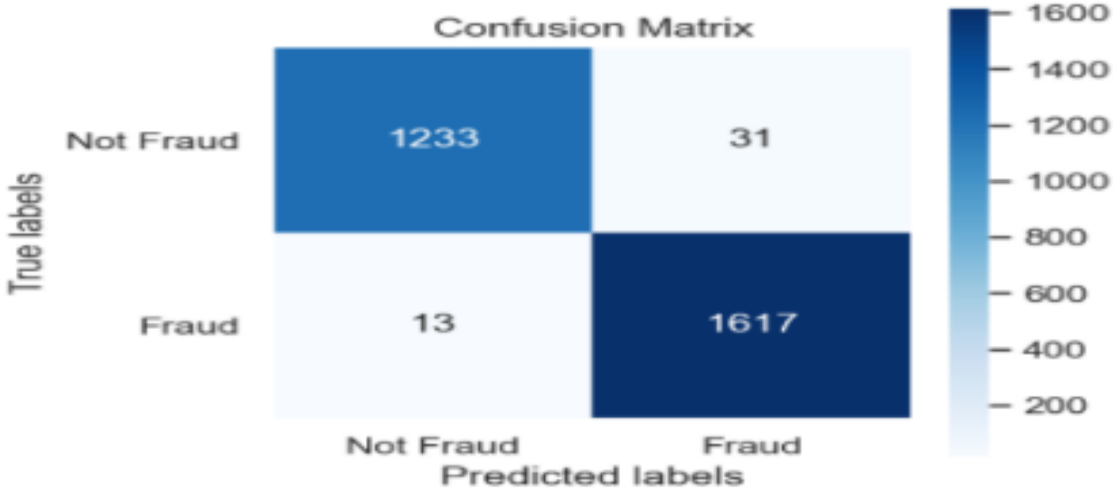
model	Test	F1	AUC	Log-Loss
Logistic Regression	using all features	0.000	0.717	0.208
	using some of the features	0.000	0.502	0.228
	using some of the features with dimensionality reduction	0.028	0.782	0.195
	using some of the features with Over-sampling using SMOTE	0.924	0.975	0.203
KNN	using some of the features	0.000	0.540	1.276
	After resampling & GridSearchCV	0.987	0.984	0.525
Random Forest	using some of the features	0.122	0.799	0.253
	using some of the features with Over-sampling using SMOTE	0.967	0.993	0.120
	using some of the features with RandomizedSearchCV & Over-sampling using SMOTE	0.964	0.992	0.152
xgboost	some of the features with Over-sampling using SMOTE	0.969	0.992	0.106
Support Vector Machine	Some of the features with Over-sampling using SMOTE	0.984	0.996	0.053

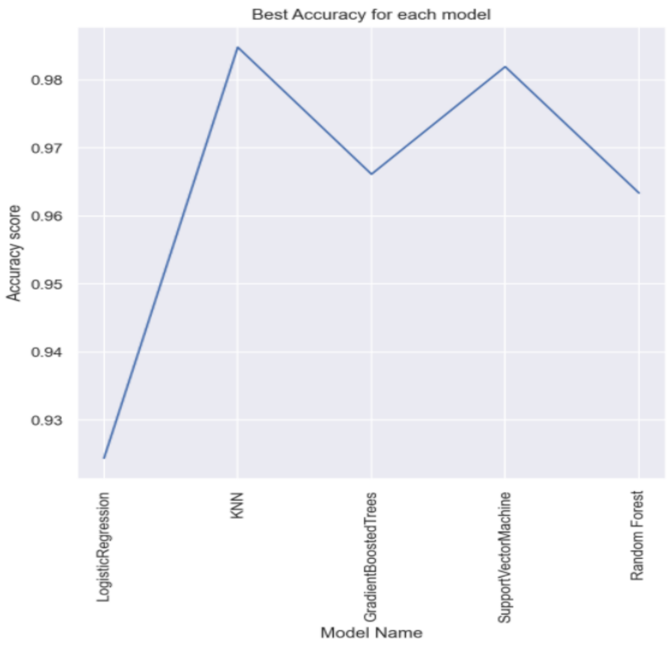
model	accuracy	f1-score	prediction	precision	recall	f1-score
Logistic Regression	0.94	0.92	Not Fraud	0.92	0.93	0.92
			Fraud	0.93	0.92	0.92
KNN	0.98	0.98	Not Fraud	0.99	0.98	0.98
			Fraud	0.98	0.99	0.99
Random Forest	0.96	0.96	Not Fraud	0.94	0.98	0.96
			Fraud	0.98	0.95	0.96
xgboost	0.97	0.96	Not Fraud	0.94	0.98	0.96
			Fraud	0.98	0.96	0.97
Support Vector Machine	0.98	0.98	Not Fraud	1.00	0.96	0.98
			Fraud	0.97	1.00	0.98

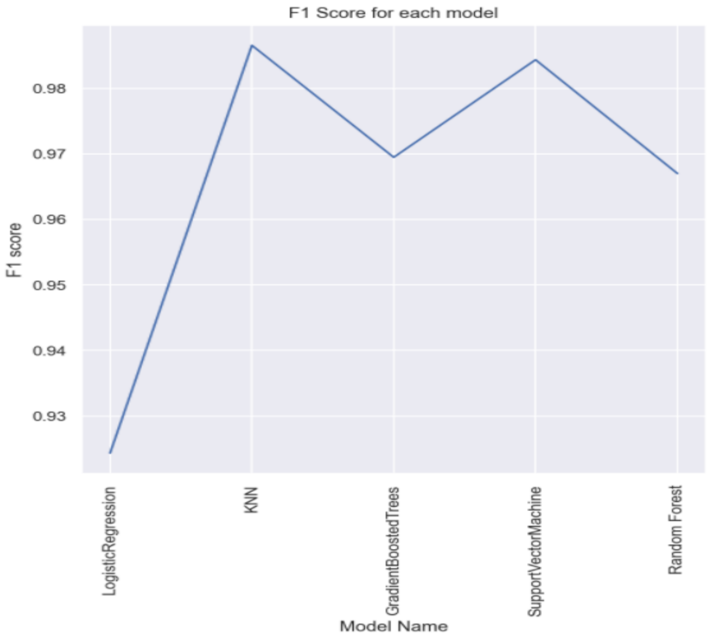
- ## Best model:

model	Test	F1	AUC	Log-Loss
KNN	After resampling & GridSearchCV	0.987	0.984	0.525

- **Graphs**







• Conclusion:

We made 5 models, We came up with the best model to do fraud detection After receiving the characteristics for each claim, the model will help insurance companies to help them detect fraud clime. We achieved very good accuracy (98%) in the best model