Belling Schrödinger's Cat to Catch the Thief

Leveraging Additional Information Embedded in Suspected but Unconfirmed Cases for Improved Predictive Accuracy in Anomaly Detection

GSrinivasaraghavan

gsr@iiitb.ac.in

Professor – Department of Computer Science, IIIT Bangalore

Partner – Performance Engineering Associates

PhD – Computer Science, IIT Kanpur

Guha.R

guha.r@wipro.com

Head – Development & GTM - Apollo Anomaly Detection Platform

Member – NASSCOM Cyber Security Taskforce

MBA – Finance, IIM Bangalore

Kartheek Palepu

kartheek.palepu@wipro.com

Associate Data Scientist – Apollo Anomaly Detection Platform

Abstract

The field of Anomaly Detection often encounters cases that are suspected, but unconfirmed for a variety of reasons. Not considering the fuzzy information embedded in these cases will lead to a loss of predictive accuracy, especially given the imbalance in the sample. This paper aims to develop a technique for leveraging information in suspected cases through an ensemble technique whereby the suspected cases have imputed labels based on Bayesian reasoning at run-time.

Introduction

Anomalies are patterns that do not conform to an expected normal behavior and finding those patterns is Anomaly Detection. Anomalies might be induced for a variety of reasons, such as malicious activity like credit card fraud, cyber-intrusion, breakdown of system but all reasons have the common characteristic that they do not conform to the normal behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains. The challenge lies in encompassing every possible normal behavior, given the boundary between normal and anomalous behavior is not precise. Include different types of Anomalies, evolving normal behavior, add multiple domains along with the prohibitively expensive effort required for availability of quality data and the task becomes that bit difficult.

Many attempts have been made to cover an exhaustive list of anomaly detection techniques, but none seem to provide a meaningful approach towards handling grey areas for significant results. This paper aims to cover that gap.

Anomaly Detection: Techniques and Application

Labels associated with data, which denotes whether an instance is normal or anomalous, to a great extent defines the anomaly detection technique applied.

Supervised Anomaly Detection approach encompasses building a predictive model for normal vs anomaly classes. Any new data instance is compared against the model to determine which class it belongs to. Though it sounds simple, it is riddled with issue of very low incidence rates arising because of imbalanced class distributions and the challenges involved with obtaining accurate and representative labels.

Semisupervised Anomaly Detection technique assumes that the training data has labeled instances only for the normal class. It involves building a model for the class corresponding to normal behavior and use the model to identify anomalies in the test data. However, obtaining a training data that covers every possible anomalous behavior that can occur in the data is difficult.

Unsupervised Anomaly Detection technique does not require any training data, warranting wider applicability. It make the implicit assumption that normal instances are far more frequent than anomalies in the test data, though it suffers from high false alarm rate if this assumption do not hold true.

The anomaly detection techniques find wide applicability in a variety of scenarios and industry domains.

- Classify Duplicates in Vendor Payments
- Unearth Claims Fraud in Auto Insurance
- Identify Discrepancies in in Media spend
- Detect Master Data Anomalies in Fixed Income Portfolio for Asset Management
- Flag High Risk Transactions from Invoices for Retailers
- Reduce Asset misappropriation and collusive Fraud
- Discover Intellectual Property Theft

Anomaly Detection: Challenges from grey areas

We refer to the doubtful cases as grey points, assuming blacks to be anomalies and whites to be normal. e.g. in fraud detection, we refer them as "suspected frauds" wherein these data points have a possibility to be either an anomaly or normal. The reasons for marking these data points in this suspected, but unconfirmed category include:

- Lack of time given investigative bandwidth
- Cost considerations and cost-benefit tradeoff
- Loss of relevant evidence
- Regulatory or user-satisfaction considerations

This brings about an interesting question around how the ‘greys’ are handled. Three trivial options come to mind:

- Ignore from analysis - this could be especially challenging where there are few, or no blacks
- Assume as white (given majority labels)- which further accentuates the imbalance

- Assume as black (play conservatively and reduce class imbalance)

All these approaches lose vital information embedded in the uncertainty. To resolve this, we draw inspiration around Erwin Schrodinger's thought experiment:

> **"***A cat is placed in a closed box with a small amount of poison. The scenario here is the cat can be either dead or still be alive or both. This is called the Principle of superposition. It states that if any object can be in one of the two states then that object is said to be in a superposition of the two.***"**

Business Scenario

Insurance claims are labelled as "Normal" and "Fraud". However, a number of cases are "suspected" claims that appear to be fraudulent to the business user but are not verified for want of time or resources (like availability of inspectors etc.)

From a technical perspective, the proportion of outlier/anomalous data has a very low incidence rate. Hence, utilizing the suspect dataset can boost the performance of the model.

Data Description

N: Normal Claims dataset where the claims are classified as genuine

F: Fraudulent Claims dataset where the claims are classified as fraudulent

G: Grey dataset claims that were suspected to be fraudulent but it could not be verified for want of time or resources

If **g** denotes any claim belonging to this set, then

$$\boldsymbol{g} \in \boldsymbol{G},\ \text{and}$$

$$\boldsymbol{g} \in \boldsymbol{G}\ (\boldsymbol{g} \in \boldsymbol{N}\ \text{OR}\ \boldsymbol{g} \in \boldsymbol{F})$$

as this data is marked as suspect by business team, chances are high that $g \in F$. This aspect is taken into consideration by assigning a suitable prior probabilities later.

Diagrammatic Representation

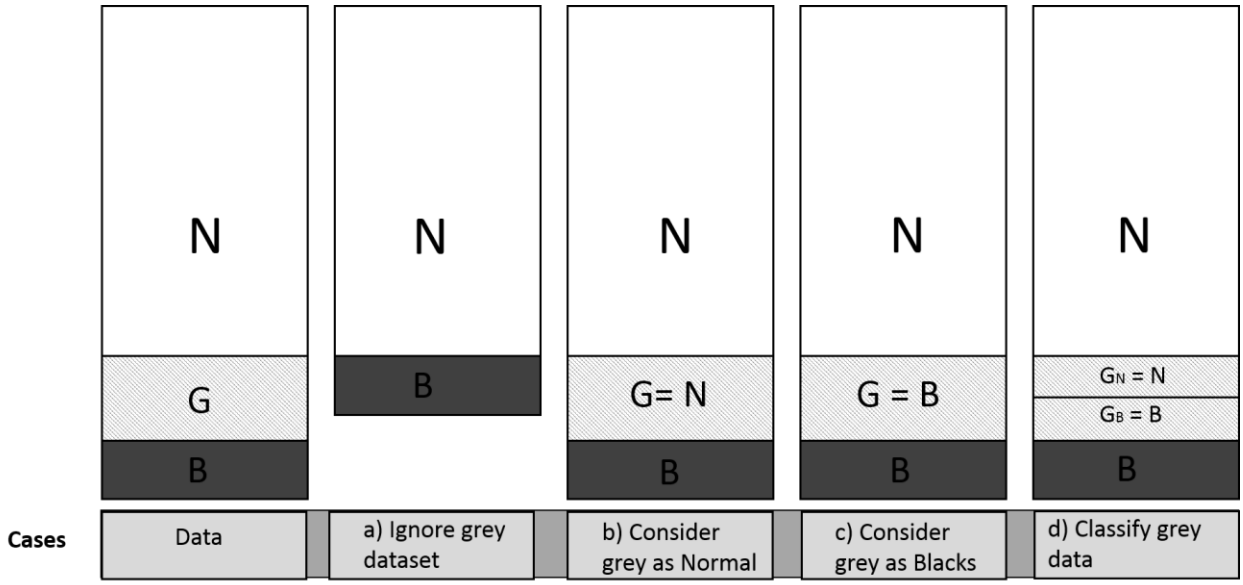


Figure – 1

	Original Data	Ignore grey dataset	Consider grey as Normal	Consider grey as Blacks	Classify grey data
Data size	$\lvert N \cup G \cup F \rvert$	$\lvert N \cup F \rvert$	$\lvert N \cup G \cup F \rvert$	$\lvert N \cup G \cup F \rvert$	$\lvert N \cup G \cup F \rvert$
Normal data size	$\lvert N \rvert$	$\lvert N \rvert$	$\lvert N \cup G \rvert$	$\lvert N \rvert$	$\lvert N \cup G_N \rvert$
Fraud data size	$\lvert F \rvert$	$\lvert F \rvert$	$\lvert F \rvert$	$\lvert G \cup F \rvert$	$\lvert F \cup G_F \rvert$

Table – 1

In Table – 1 **|.|** represents the size of the data

Fig. 1 shows the different options

a) Ignore grey records
b) Consider grey records as normal claims
c) Consider grey records as fraudulent claims
d) Classify grey data as normal/fraud before building the model

Objective

The objective of this exercise is to investigate if

a) The model created by the classification of each element of set G (case d above):

$$\forall\, \boldsymbol{g} \in \boldsymbol{G}\ (\boldsymbol{g} \in \boldsymbol{N}\ \text{OR}\ \boldsymbol{g} \in \boldsymbol{F})$$

Results in a better model compared to the models that either ignore set G or consider set G as part of set N or set F (see diagram 1)

b) The above leads to discovery of new patterns for fraud detection

Datasets Overview

Three different data sets have been used. In all datasets, a small proportion of claims are marked as **greys** (suspected frauds), **known frauds** and others as **normal**. The summary of the datasets are given below:

- **Dataset – 1:** This dataset pertains to Vehicle insurance claims. The reasons for greys include trade-off around cost/savings of an investigation, lack of evidence, process failures etc.
- **Dataset – 2:** This dataset pertains to financial payments for a large business house where exercise was to determine potential duplicates. Reasons for greys include limited time-window to complete investigation, limited background documents etc.
- **Dataset – 3:** This dataset pertains to financial bonds from a large business house where exercise was to determine potential anomalies. Reasons for greys include limited background documents.

The following Table – 2 gives relevant statistics about the datasets:

	Dataset – 1	Dataset – 2	Dataset - 3
Number of Data points	8,709	52,087	56,169
Number of Attributes	59	15	50
Categorical Attributes	8	0	2
Number of Normal' s (Whites)	8,627	51,774	55,357
Frauds Confirmed (Blacks)	31	131	379
Frauds Suspected (Greys)	51	182	433
Incident Rate (Blacks)	0.35 %	0.25 %	0.67%
Grey Incident Rate	0.58 %	0.34 %	0.77%

Table – 2, Datasets Description

Problem Formulation

Performance definition: Performance of the model is defined a measure of its ability to correctly classify new data as normal/fraud.

Let

$\boldsymbol{P_{NI}}$: Denote the performance of the model (in classifying new data as normal/fraud) when set G is ignored

$\boldsymbol{P_{NB}}$: Denote the performance of the model when all elements of set G are assumed as fraud (Black)

$\boldsymbol{P_{NW}}$: Denote the performance of the model when all elements of set G are assumed as normal (White)

$\boldsymbol{P_{NC}}$: Denote the performance of the model built with data in set G classified as normal/fraud

Then,

$$\boldsymbol{P_{NC} > max(P_{NI}, P_{NB}, P_{NW})}$$

i.e. the model built by incorporating the classification of the grey set G performs better as compared to the other models.

Approach Outline

a) We can use any two class classifiers for normal and fraudulent claims. With this model, we can evaluate $\boldsymbol{P_{NI}, P_{NB}, P_{NW}}$ as above.

b) Use Naïve Bayes Classifier to label the dataset $\boldsymbol{G}\ as$

$\boldsymbol{G_N}$ for all records

$$\boldsymbol{g: g \in G \ \wedge \ g \in N} \quad \text{and}$$

$\boldsymbol{G_F}$ for all records

$$\boldsymbol{g: g \in G \ \wedge \ g \in F} \quad \text{(See fig above)}$$

c) With this classification, we can compute P_{NC}

d) With the data, we can show that

$$\boldsymbol{P_{NC} > max(P_{NI}, P_{NB}, P_{NW})}$$

Naïve Bayes

Definition:

Conditional Independence: Let X, Y, Z be random variables. If variables X, Y are independent.

Given Z, then X, Y are said to be conditionally independent. It is denoted as

$$\boldsymbol{X \perp\!\!\!\perp Y \mid Z} \quad \boldsymbol{\ldots\ldots\ldots\ldots(1)}$$

We can also express the joint distribution of X and Y given Z as

$$\boldsymbol{P(X,\ Y \mid Z) = P(X \mid Z)\,.\,P(Y \mid Z)} \quad \boldsymbol{\ldots\ldots\ldots\ldots\ (2)}$$

Assumption:

Given Labels (Y), the attributes (X) of any dataset are conditionally independent i.e.

$$\boldsymbol{X_i \perp\!\!\!\perp X_j \mid Y_1} \quad for\ all\ i\ \&\ j\ and\ i \neq j$$

Using (**2**) above, we can write the joint distribution of data points as equation(**3**):

$$\boldsymbol{P(X_1, X_2, \ldots\ldots, X_n|Y_1)} = \boldsymbol{P(X_1|Y_1).P(X_2|Y_1)\ldots\ldots P(X_n|Y_1)}$$

$$= \boldsymbol{\prod_{i=1}^{n} P(X_i|Y_1)} \quad \ldots\ldots\ldots\ldots\ (\mathbf{3})$$

Interpretation of Insurance Data

Attributes	$\mathbf{X_1}$	$\mathbf{X_2}$	.	.	.	$\mathbf{X_n}$	**Classification(Y)**
Data types	R	C	R	R	C	R	C
							Normal
							Normal
							.
			.	.	.		.
							.
							Fraud
							Fraud

Table – 3: Structure of insurance claims data with Classification

We can interpret the insurance claims data in Table – 2 as:

a) Column – Classification(Y): Is Normal/Fraud is the label.
b) Columns – X_1, X_2, ..., X_n: are the attributes that are conditionally independent given the label.

Naive Bayes Classification of Grey dataset G

Two models are first built one for Normal data set & and another for Fraud data set. Grey data are then classified using these models.

The insurance data has both numeric and categorical attributes.

The approach taken for the numeric variables is to assume them (each column) to have a Gaussian distribution $N(\mu, \sigma^2)$. Parameters for each numeric attribute column is estimated.

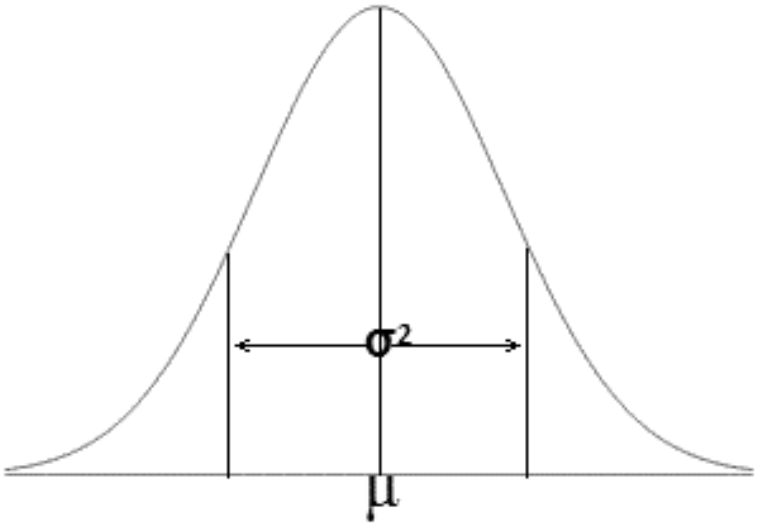


Fig – 2: Gaussian distribution for the data in the Column for a given classified $N(\mu,\sigma^2)$

Attribute	X_1	X_2	..	..	..	X_n	Y
Data type	C	N	N	C	C	N	C

Fig – 3: Gaussian distribution for a numeric attribute

C : Categorical Data , N : Numerical Data

For the Categorical data, the proportion is calculated for each possible value in that column for a given label.

For classifying any grey data, now we can calculate the probability of each attribute for a given classification using the Naive Bayes model.

For the categorical attributes, we use Laplacian smoothing, if that value doesn't exist in the training data set.

$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$	..	..	..	$\mathbf{P_n}$

Fig – 4: shows the probability of each attribute of a grey data.

Let $P_1, P_2, P_3, \dots, P_n$ be the probabilities of each attribute of a grey data. This is calculated using the corresponding proportion/Gaussian distribution of the model.

As the attributes are conditionally independent, we can arrive at the joint distribution of the grey data as equation($\mathbf{4}$):

$$\boldsymbol{P\left(D_{grey}\middle|Label\right)} = \boldsymbol{P_1 . P_2 \dots . P_n}$$

$$= \boldsymbol{\prod_{i=1}^{n} P_i} \qquad \boldsymbol{\dots\dots\dots\dots\dots(4)}$$

Where:

$\boldsymbol{D_{grey}}$ is the grey data element/record and Label can be B or W for fraud & Normal

Given the above likelihood of the data, we can get the posterior for the Label using a suitable prior probabilities using Bayes theorem:

$$\boldsymbol{P(W\ |D_{grey})\ \alpha\ P\left(D_{grey}\middle|W\right).P(W)} \ \boldsymbol{\dots\dots\dots\dots\dots}\ (\mathbf{5})$$

$$\boldsymbol{P(B\ |D_{grey})\ \alpha\ P\left(D_{grey}\middle|B\right).P(B)} \ \boldsymbol{\dots\dots\dots\dots\dots}\ (\mathbf{6})$$

Where:

$\boldsymbol{P(W)}\ and\ \boldsymbol{P(B)}$ are prior probabilities

As the business team "Suspects" the grey set G, as fraud claims, we choose the following values for the probabilities

$$\boldsymbol{P(B) = 0.65} \ \textit{and} \ \boldsymbol{P(W) = 0.35} \ \ldots\ldots\ldots\ldots\ (7)$$

These values are subjective and have been chosen in consultation with the domain experts.

We can then express the ratio of the Posterior γ as

$$\boldsymbol{\gamma} = \frac{\boldsymbol{P(B|D_{grey})}}{\boldsymbol{P(W|D_{grey})}}$$

$$\boldsymbol{\gamma} = \frac{\boldsymbol{P\left(D_{grey}|B\right).P(B)}}{\boldsymbol{P\left(D_{grey}|W\right).P(W)}} \ \ldots\ldots\ldots\ldots\ (8)$$

Using the values from (7), we frame equation (8) as

$$\boldsymbol{\gamma} = \frac{\boldsymbol{P\left(D_{grey}|B\right).(0.65)}}{\boldsymbol{P\left(D_{grey}|W\right).(0.35)}}$$

We now apply the following criteria to identify each claim g from the grey set G as:

$$\boldsymbol{Classification\ of\ g} = \begin{cases} \boldsymbol{Black}, & \boldsymbol{\gamma \geq 1} \\ \boldsymbol{White}, & \boldsymbol{\gamma < 1} \end{cases}$$

Once the elements of set G are classified, we create the final model with the classified data (see figure - 1: Case (d)). With this model, we can calculate its performance P_{NC} on new dataset. We see from the data that:

$$\boldsymbol{P_{NC} > max(P_{NI}, P_{NB}, P_{NW})}$$

Exercised Technique

Whenever an unconfirmed case is encountered we have neglected those cases or sometimes treated them as anomalies and started building the model. But, in this paper once the dataset with unconfirmed cases is encountered, we change the unconfirmed cases into either whites or blacks using equation (8) i.e. we first build a model for the data with whites and then build a model for the data with blacks. Using these two models we calculate the value of γ for each grey point which in turn is used to classify the grey points into whites or blacks. Once the grey points are classified we build a model using the overall data including the grey data points.

Interpretations

1. We have built models for all the cases in Figure – 1 using all the 3 described datasets in Table – 2. We have plotted the F**5** Score (a trade-off score between Recall and Precision) for each of these datasets.

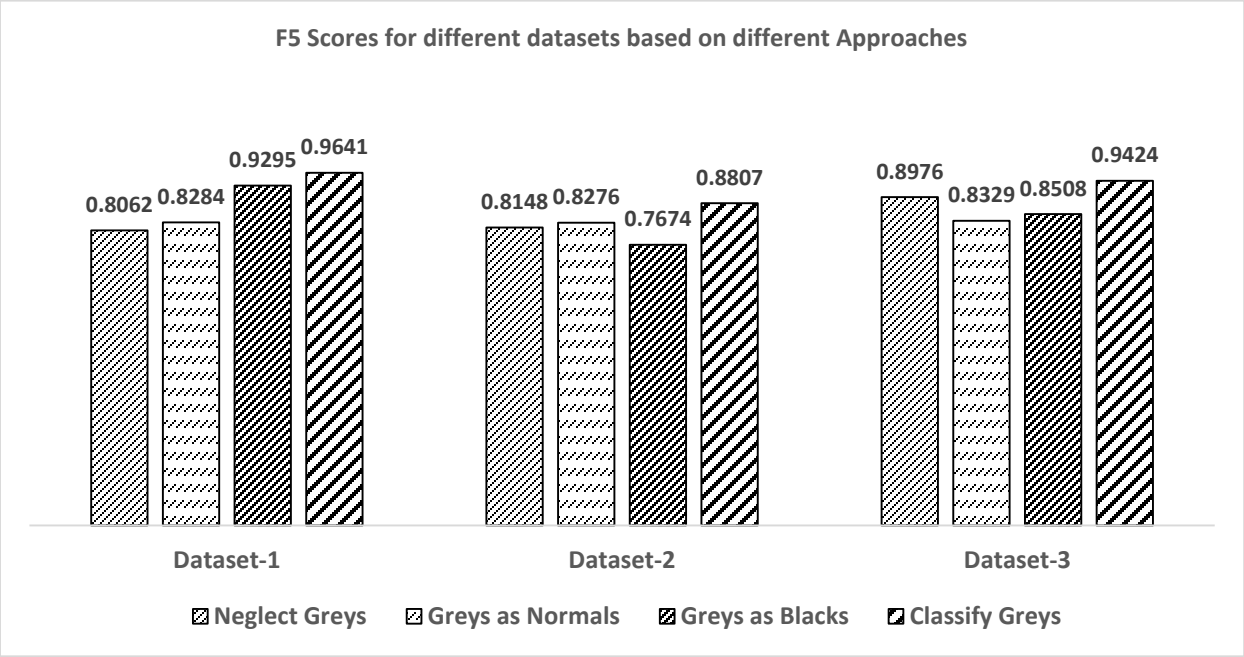


Figure –5

We can observe that there is a significant boost in the performance in each of the datasets when we started classifying the greys to build a model.

2. We have used the Dataset – 2 to generate 15 synthetic datasets by picking random points from blacks and some random points from white which are finally marked as greys and then the classification is done over it. The datasets description along with the resultant F5 Scores are presented below.

	Number of Data points	Number of Attributes	Number of Normal' s (Whites)	Frauds Confirmed (Blacks)	Frauds Suspected (Greys)	Incident Rate (Blacks)	Grey Incident Rate
SD-1	52087	14	50056	224	1807	0.43	3.46
SD-2	52087	14	50942	196	949	0.37	1.82
SD-3	52087	14	50370	174	1543	0.334056	2.962351
SD-4	52087	14	51141	209	737	0.401252	1.41494

SD-5	52087	14	51150	217	720	0.416611	1.382303
SD-6	52087	14	50403	174	1510	0.334056	2.898996
SD-7	52087	14	51343	209	535	0.401252	1.027128
SD-8	52087	14	50687	174	1226	0.334056	2.353754
SD-9	52087	14	49370	224	2493	0.43005	4.786223
SD-10	52087	14	50301	209	1577	0.401252	3.027627
SD-11	52087	14	51454	196	437	0.376294	0.838981
SD-12	52087	14	50153	209	1725	0.401252	3.311767
SD-13	52087	14	51530	196	361	0.376294	0.693071
SD-14	52087	14	51428	131	528	0.251502	1.013689
SD-15	52087	14	51660	174	253	0.334056	0.485726

Table – 4, Datasets Description

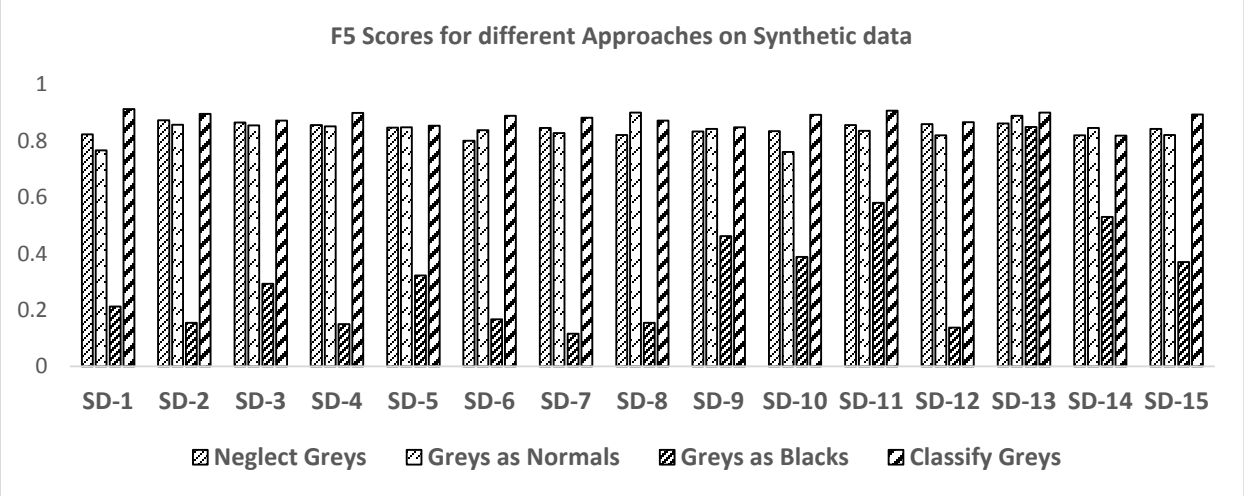
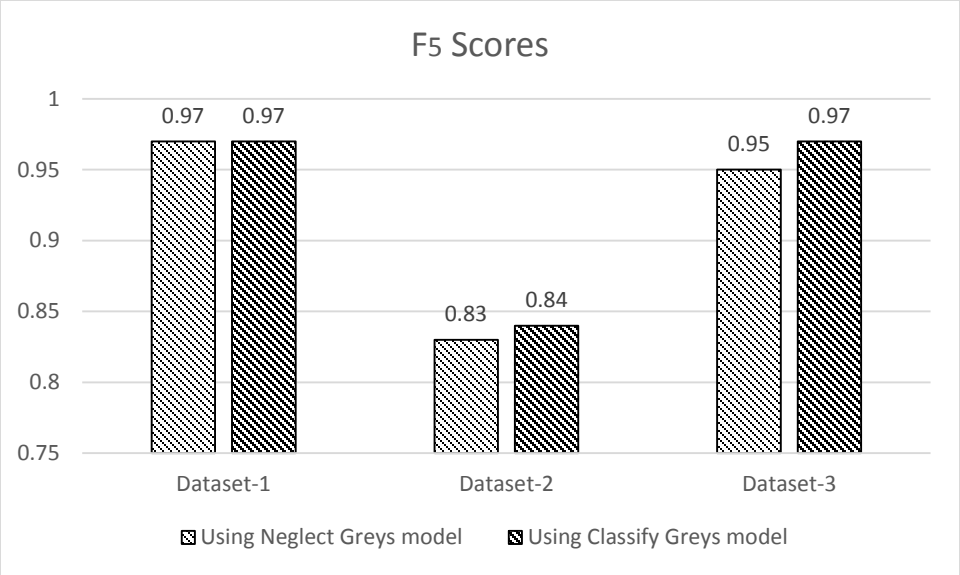


Figure –6, Resultant F5 Scores

In case of 13 out of 15 synthetic datasets we can observe that there is a significant boost in the performance when we started classifying the greys to build a model.

3. We have also double checked the approach by applying the Neglect Greys model and Classify Greys model on Neglected Greys data. The hypothesis is that the Classify Greys model should perform better than the Neglect Greys model.



Key Finding

Based on the exercise, we came up with the following finding:

— *There is a significant value in tapping the intelligence embedded in 'greys' data points – works better than completing ignoring them or marking them as either white or black.*

References

1. *Anomaly detection: A survey, Chandola et al, ACM Computing Surveys (CSUR), Volume 41 Issue 3, July 2009.*
2. *Minority Report in Fraud Detection: Classication of Skewed Data – by Clifton Phua, Damminda Alahakoon, and Vincent Lee. Sigkdd Explorations, Volume – 6, Issue – 1.*
3. *Fawcett T and Provost F. "Adaptive fraud detection", Data Mining and Knowledge Discovery, Kluwer, 1, pp 291-316, 1997.*
4. *Bayesian Reasoning and Machine Learning – by Barber. D, Published by Cambridge University, 2012.*
5. *Make Decisions with Sequential Observations: A Simulation – by Theodore W. Frick.*
6. *Ensemble Methods in Machine Learning – by Thomas G Dietterich, Oregon State University Corvallis Oregon USA.*
7. *Detecting Auto Insurance Fraud by Data Mining Techniques – by Rekha Bhowmik, Computer Science Department, University of Texas at Dallas, USA, International Journal of Computer Applications (0975 - 8887), Volume 79 – No.2, October 2013*
8. *Machine Learning Techniques for Anomaly Detection: An Overview – by Salima Omar, Asri Ngadi and Hamid H. Jebur, International Journal of Computer Applications (0975 - 8887), Volume 79 – No.2, October 2013.*
9. *A Classification Framework for Anomaly Detection – by Ingo Steinwart, Don Hush and Clint Scovel, Journal of Machine Learning Research 6 (2005) 211–232.*
10. *Comparative Analysis Of Machine Learning Techniques For Detecting Insurance Claims Fraud, Wipro Technologies – by R Guha, Kartheek Palepu and Shreya Manjunath.*