Exploring the role of Big-data in finance domains.

Abhishek Shastry HM(A20468925)

CSP 554 Big Data Technologies

Illinois Institute of Technology

ashastry1@hawk.iit.edu

Abstract:

The below research paper focuses on the usage of big data in finance. The different finance fields constitutes stock exchange, banking institution and investment strategies.
We will see how volume, variety, velocity and veracity effects the overall performance,

This paper will compare both structured and unstructured in financial domains. We will also look on the disadvantages and advantages of using big data in finance. We will look some recent and live examples out in the market. At last this paper will focus on the current challenges in the field, scope of improvement left, future possible occurrences.
Keywords: Unstructured, finance, analysis, decision.

I. Introduction:

Big data is widely used to take analytical decision at finance industry, to implement better investment decisions with consistent returns, maximize the portfolio returns and help to take better trading decisions. The sentiments of the specific topics of the company are incorporated into the stock prediction model. NLP can be used here as a helping hand to parse the unstructured messages from the social media and extract its features for the price prediction. Many companies lack responsive digital FP&A(financial planning and analysis) function built on integrated technologies with flexible operating models for rapid decision making.

Big data alone will not help in driving the decision/forecasting, it makes complete sense when integrated with AI. Financial institutions have long stored unanalyzed data. The data in the finance industry keeps continuously changing and hence the speed at which information is compiled and decisions are made needs to be shortened to the point where these activities often occur

in near real time. Extracting value from uncertain data can be a challenging task in banking and financial markets. Banking sectors are looking to transform their approach from product-centric to customer-centric organizations. Being a customer centric helps to understand their customers better and build manful relationships.

Technologies such as Hadoop/mapreduce and Nosql are among least implemented in banking and financial sectors. Among the multiple big data sources, transactions and log data have been used frequently whereas social media , audio, video have been under utilized. Datasets are often too large for business or data analysts to view and analyze with traditional reporting and data mining tools thereby increasing the need of more advanced data visualization and analytics capabilities. Big data have the capability to analyze the most complex data unstructured such as the texts, voice(call center analysis and its feedback), video(customer behavior), geospatial(helps to penetrate the market and beat the competitors, execute campaigns)which can be used to derive the hidden insights for better decision.

Some of the key fields and applications of big data analytics in finance include real time stock markets, risk analysis, increased revenue generation, automating operational pipeline , generating reports, cloud based integration Challenges of big data as related to finance are a bit more complex due to many reasons such as regulatory acts of government laws, data security, privacy, miss use of data etc.

Many companies believe that use of big data benefits to achieve the goal but are doubt-full on adopting the technologies. Many firms who have opted are releasing the difference through it. Problems existing in the current market are some of the prior reasons for why adopting makes complete sense and here's some.

Examples in the current market:

1) https://www.bloomberg.com/news/articles/2021-06-27/a-27-billion-pile-of-debt-looms-over-india-s-new-bad-bank

Bad loans is one of the most frequent and worst scenarios which puts countries economy in trouble. Banks have failed to recover on time and blind faith on some of the giants. "People lie but data never lies"

2) https://www.worldbank.org/en/topic/competitiveness/publication/global-investment-competitiveness-report

Big investors fail to recognize the opportunities due to lack of visibility and supportive evidences. Big data can combine by adding up evidences, strategies, futuristic outcomes such as risks, advantages by considering all factors which was once considered redundant. These factors can be considered as features in driving the decisions.

3) https://www.investmentmonitor.ai/finance/how-digitisation-is-transforming-financial-services

Many developing countries have now started the use of digital payments, banking etc. The rise of such vast amount of data requires a need of strong technology based support. India saw a rise of 9.69 billion transactions from 5.93 billion in the previous year in terms of volume.

2. Magic of R squared stocks and returns:

R squared is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable in a regression model. R Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variables in a regression model.
In investing, R-squared is generally interpreted as the percentage of a fund or security's movements that can be explained by movements in a benchmark index.
An R-squared of 100% means that all movements of a security i.e other dependent variables are completely explained by movements in the index i.e the independent variables(Y=security's price movement , x1,x2…xn =index movements)An R squared value between 70 to 100 tells that there is a strong correlation between returns of the portfolio and the bench mark index.

An Adjusted R square comes into picture when multiple predictor variables comes into picture. Thus, a model with more terms may seem to have a better fit just for the fact that it has more terms, while the adjusted R squared compensates for the addition of variables and only increases if the new term enhances the model.

n [13]:
```
pdf_financial_banking_data=pd.read_table("/Users/abhishekshastry/Downloads/table.txt")
pdf_financial_banking_data
```

ut[13]:

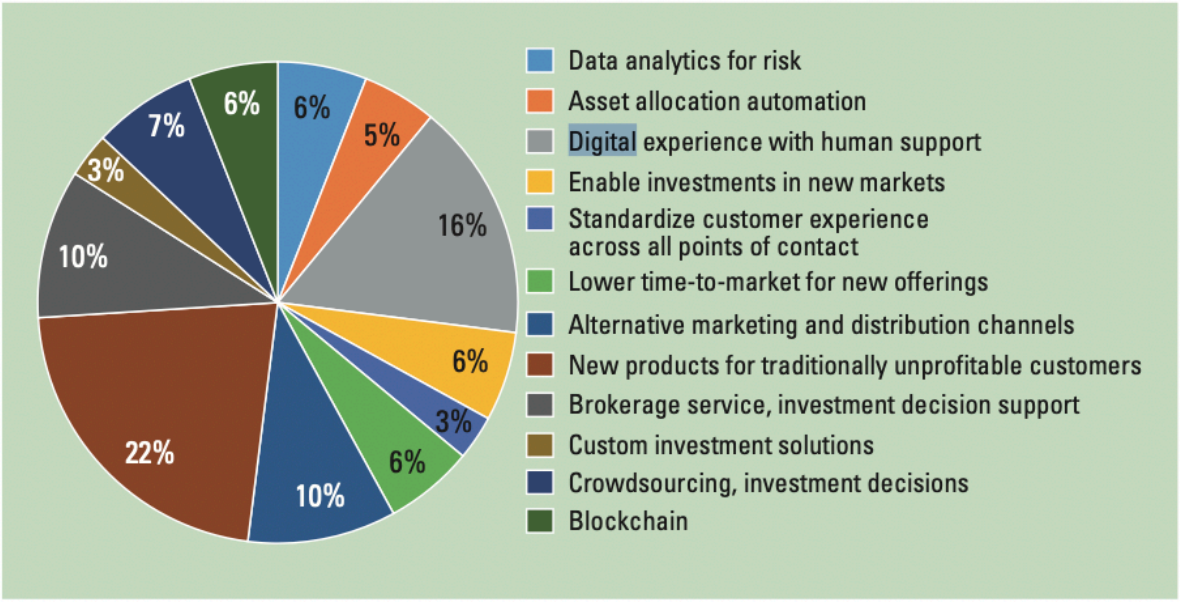
	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
0	18.9	350.0	165	260	8.00	2.56	4	3	200.3	69.9	3910	1
1	17.0	350.0	170	275	8.50	2.56	4	3	199.6	72.9	3860	1
2	20.0	250.0	105	185	8.25	2.73	1	3	196.7	72.2	3510	1
3	18.3	351.0	143	255	8.00	3.00	2	3	199.9	74.0	3890	1
4	20.1	225.0	95	170	8.40	2.76	1	3	194.1	71.8	3365	0
5	11.2	440.0	215	330	8.20	2.88	4	3	184.5	69.0	4215	1
6	22.1	231.0	110	175	8.00	2.56	2	3	179.3	65.4	3020	1
7	21.5	262.0	110	200	8.50	2.56	2	3	179.3	65.4	3180	1
8	34.7	89.7	70	81	8.20	3.90	2	4	155.7	64.0	1905	0
9	30.4	96.9	75	83	9.00	4.30	2	5	165.2	65.0	2320	0
10	16.5	350.0	155	250	8.50	3.08	4	3	195.4	74.4	3885	1
11	36.5	85.3	80	83	8.50	3.89	2	4	160.6	62.2	2009	0
12	21.5	171.0	109	146	8.20	3.22	2	4	170.4	66.9	2655	0
13	19.7	258.0	110	195	8.00	3.08	1	3	171.5	77.0	3375	1
14	20.3	140.0	83	109	8.40	3.40	2	4	168.8	69.4	2700	0
15	17.8	302.0	129	220	8.00	3.00	2	3	199.9	74.0	3890	1
16	14.4	500.0	190	360	8.50	2.73	4	3	224.1	79.8	5290	1

```
29  16.5  350.0  165  255  8.50  2.73    4    3  185.2  69.0   3660       1
```

```
]: regression_model = LinearRegression()

# Fit the data(train the model)
regression_model.fit(pdf_financial_banking_data.iloc[:,1:12],pdf_financial_banking_data.iloc[:,0:1])

# Predict
y_predicted = regression_model.predict(pdf_financial_banking_data.iloc[:,1:12])

# model evaluation
mse=mean_squared_error(pdf_financial_banking_data.iloc[:,0:1],y_predicted)

rmse = np.sqrt(mean_squared_error(pdf_financial_banking_data.iloc[:,0:1], y_predicted))
r2 = r2_score(pdf_financial_banking_data.iloc[:,0:1], y_predicted)

# printing values
print('Slope:' ,regression_model.coef_)
print('Intercept:', regression_model.intercept_)
print('MSE:',mse)
print('Root mean squared error: ', rmse)
print('R2 score: ', r2)
```

```
Slope: [[-7.79461371e-02 -7.33986565e-02  1.21114874e-01  1.32903370e+00
   5.97598884e+00  3.04177907e-01 -3.19857619e+00  1.85362227e-01
  -3.99146171e-01 -5.19333628e-03  5.98654555e-01]]
Intercept: [17.77320377]
MSE: 6.245598443464639
Root mean squared error:  2.4991195336487286
R2 score:  0.835271314016021
```

Here R square value of 83.5 percent of variation of dependent variable is explained by independent variables.



Approximate proportion of the FinTech market by area.

3.Credit Card Fraud detection

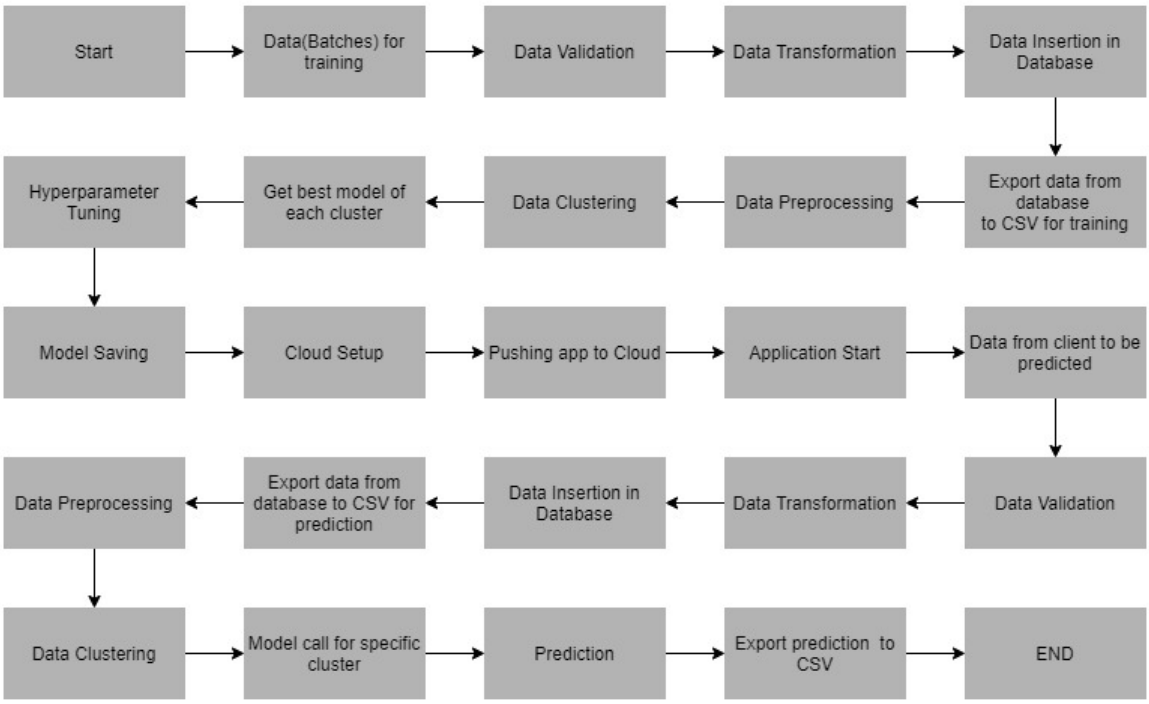
There are many metrics that needs to be considered while creating the feature columns for predicting the banking/financial market as there are many regulations that comes each and every year and it varies and a constant update needs to be done in order to stay in the current trend these include XVA (valuation adjustments of derivative instruments, based on counterparty credit risk, cost of funding, margin, and so on).

I am sequenced and transactional data can be used to analyze the consumer and behavioral market in the bank. These two metrics are often found with every banks and hence can be utilized to achieve the goal. One of the common thing that I noticed is that by analyzing the trade volume with respect to the time they say the default loans can be analyzed but but I disagree as sometimes variation in the market trade can go very low due to unexpected situations like covid, recession, .com crash etc and that cannot be used as a critical value for the judgment of the defaulters loan.

Many large financial institutions consider that big data algorithms can definitely solve the technical problems but not the business problems. But they need to be made aware that the big data algorithms have the capability to see all the business problems as well as not just only the technical related problems in the institution.
In traditional Banking systems the banks collect the information of the user from the applicant themselves and also from the various external sources. The external sources include some of the financial institutions and the non-financial sector who provide services online for the credit score verification. These credit scores are taken from the banks to judge and provide a value and an interval to the credit that can be given to the applicant but what they mess out here is that these centralized database is they do not update frequently and they may update once in a week or once a month providing a lagging system in the current trend of the applicant.
Here I see the space where the big data technology can bounce into the picture here for every single data point or a transaction linear regression can we plotted with probability of default and the various other transactions.
Below is the credit card fraud detection algorithm which can be utilized by

various firms.

Architecture To build a classification methodology to determine whether a person defaults the credit card payment for the next month.



Sample code for training the ML model for fraud detection.

Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [ ]:

```
# Doing the necessary imports
from sklearn.model_selection import train_test_split
from data_ingestion import data_loader
from data_preprocessing import preprocessing
from data_preprocessing import clustering
from best_model_finder import tuner
from file_operations import file_methods
from application_logging import logger
import numpy as np
import pandas as pd
class trainModel:
    def __init__(self):
        self.log_writer = logger.App_Logger()
        self.file_object = open("Training_Logs/ModelTrainingLog.txt", 'a+')
    def trainingModel(self):
        # Logging the start of Training
        self.log_writer.log(self.file_object, 'Start of Training')
        try:
            # Getting the data from the source
            data_getter=data_loader.Data_Getter(self.file_object,self.log_writer)
            data=data_getter.get_data()
            preprocessor=preprocessing.Preprocessor(self.file_object,self.log_writer)
            X,Y=preprocessor.separate_label_feature(data,label_column_name='default payment next month')
            # check if missing values are present in the dataset
            is_null_present,cols_with_missing_values=preprocessor.is_null_present(X)
            # if missing values are there, replace them appropriately.
            if(is_null_present):
                X=preprocessor.impute_missing_values(X,cols_with_missing_values)
            kmeans=clustering.KMeansClustering(self.file_object,self.log_writer) # object initialization.
            number_of_clusters=kmeans.elbow_plot(X)  #  using the elbow plot to find the number of optimum clusters
            # Divide the data into clusters
            X=kmeans.create_clusters(X,number_of_clusters)
            #create a new column in the dataset consisting of the corresponding cluster assignments.
            X['Labels']=Y
            # getting the unique clusters from our dataset
            list_of_clusters=X['Cluster'].unique()

            """parsing all the clusters and looking for the best ML algorithm to fit on individual cluster"""

            for i in list_of_clusters:
```

```
        for i in list_of_clusters:
            cluster_data=X[X['Cluster']==i] # filter the data for one cluster

            # Prepare the feature and Label columns
            cluster_features=cluster_data.drop(['Labels','Cluster'],axis=1)
            cluster_label= cluster_data['Labels']

            # splitting the data into training and test set for each cluster one by one
            x_train, x_test, y_train, y_test = train_test_split(cluster_features, cluster_label, test_size=1 / 3
            # Proceeding with more data pre-processing steps
            train_x = preprocessor.scale_numerical_columns(x_train)
            test_x = preprocessor.scale_numerical_columns(x_test)

            model_finder=tuner.Model_Finder(self.file_object,self.log_writer) # object initialization

            #getting the best model for each of the clusters
            best_model_name,best_model=model_finder.get_best_model(train_x,y_train,test_x,y_test)

            #saving the best model to the directory.
            file_op = file_methods.File_Operation(self.file_object,self.log_writer)
            save_model=file_op.save_model(best_model,best_model_name+str(i))

        # logging the successful Training
        self.log_writer.log(self.file_object, 'Successful End of Training')
        self.file_object.close()

    except Exception as e:
        # logging the unsuccessful Training
        self.log_writer.log(self.file_object, 'Unsuccessful End of Training')
        self.file_object.close()
        raise Exception
```

Interesting fact: (One of the strong proof of data why we need big data!!)

In 2016, one million links were shared, two million friend requests were made and three million messages were sent every 20 minutes on Facebook.
1,540,000,000 users active at least once a month
974,000,000 smartphone users;
12% growth in users between 2014 and 2015;
81 million Facebook profiles
20 million applications installed on Facebook every day.

Usage statistics for big data tools according to a survey of 2,895 respondents from the data analytics community and vendors.

Tool	2016	2015	2015 –> 2016
Hadoop	22.1%	18.4%	+20.5%
Spark	21.6%	11.3%	+91%
Hive	12.4%	10.2%	+21.3%
MLlib	11.6%	3.3%	+253%
SQL on Hadoop tools	7.3%	7.2%	+1.6%
H2O	6.7%	2.0%	+234%
HBase	5.5%	4.6%	+18.6%
Apache Pig	4.6%	5.4%	–16.1%
Apache Mahout	2.6%	2.8%	–7.2%
Dato	2.4%	0.5%	+338%
Datameer	0.4%	0.9%	–52.3%
Other Hadoop/HDFS-based tools	4.9%	4.5%	+7.5%

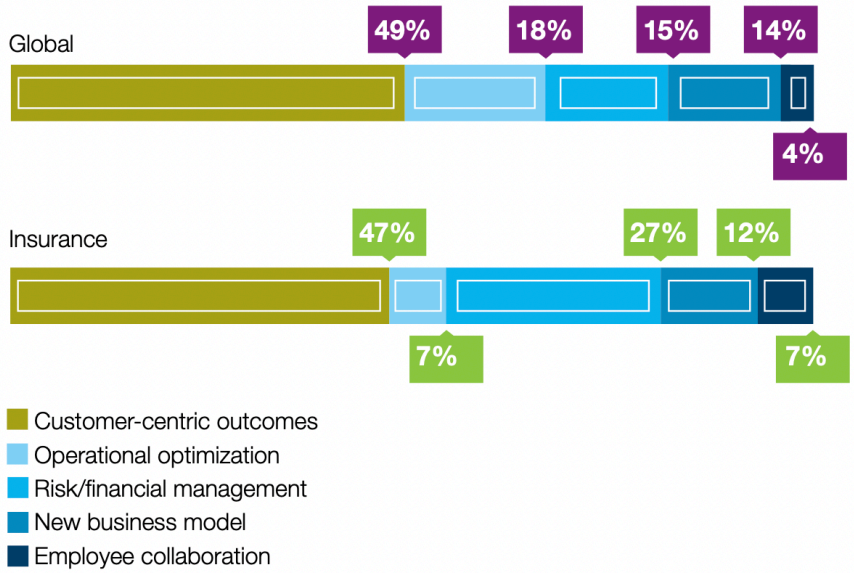
4.Role of Big data in Insurance:

Data are at the heart of the relationship between policyholders and insurers and help in establishing the conditions for their mutual commitment. Insurers conduct their business and manage the risks. Knowledge of risk and client behavior is the key component in ensuring the sustainability of insurance business. Identifying the requirement of the people and the risk calculation becomes very crucial role in finding out the insurance opportunities. Not all the countries in the world they use insurance up to its full capacity insurance policies are many unknown and also large number of populations are not covered including the health insurance , the life insurance, house insurance etc.

The data that is collected for the financial sector analysis are from a wide variety of sources. This can be broadly classified into structured and structured and semi structured. Some of the structured data sources are trading systems, account systems, securities reference data. Price information, technical indicators. Some of the unstructured data sources include daily stock feeds, online news , customer feedback, articles, announcements. Some of semi structured data sources include Financial products Markup Language, Market Data Definition Language etc. For banks and financial services providers, the volume of data they generate, consume, store, and access will increase exponentially year over year. The quality in the data selection plays a significant role here.For each requirement in the sector, this section presents applicable technologies and the research questions to be developed.

Technical Requirement	Technology	Research Question
Data Acquisition	Acquisition pipeline	Data stream management
	APIs technology	Privacy and anonymization at collection time
		Social APIs
Data quality	Manual processing and validation	Scalable data curation and validation
		New methods to improve precision and reliability
Data extraction	Language modelling	Statistical language models
	Machine Learning	Required inference functionality
	Scalability in real-time	Processing of large datasets
Data integration / sharing	Wrappers/mediators to encapsulate distributed & automatic data and schema mapping	User-specific integration
		Data variety: sentiments, quantitative information
		Scaling methods for large data volumes and near-real time processing.
Decision support	Multi-attribute decision models	Stream-based data mining
		Machine learning adaptation to evolving content
	Resource allocation in mining data streams	Improved storage, computation and communication capabilities
Data privacy & security	Roles-based IdM and access control	Privacy by design | Security by design
		Data Security for public-private hybrid environments
		Enhanced Compliance management
	Database encryption NoSQL	Apply external encryption and authentication controls

Big data objectives



Half of the big data efforts underway by insurance companies are focused on achieving customer-centric outcomes.

The 7 important ways in which big data is helping the insurance industry are: 1. Customer Acquisition, 2. Customer Retention, 3. Risk Assessment, 4. Fraud Prevention and Detection, 5. Cost Reductions, 6. Personalized Service and Pricing, 7. Effects on internal processes. Big data has accounted for more than 2.4B$ of investment in insurance industry alone in 2018.

https://www.reportbuyer.com/product/5482376/big-data-in-the-insurance-industry-opportunities-challenges-strategies-forecasts.html?utm_source=datafloq&utm_medium=ref&utm_campaign=datafloq

In a study conducted by EIOPA(European Insurance and Occupational Pensions Authority) found that the most significant role of big data in insurance today is in pricing and underwriting. The study shows a fascinating data which is I totally agree. One of the most interesting uses of big data is when it is used as a tool to predict and even change customer behavior. This is tied into the IoT; insurers who can correctly analyze customer behaviors using data from a wide range of devices may be able to step in before a claim is even made to remind policyholders to adjust high-risk behaviors, such as driving too fast or forgetting to set a burglar alarm.The role of big data in fraud detection is also significant. Reports estimate about 1300 insurance scams are detected every day.

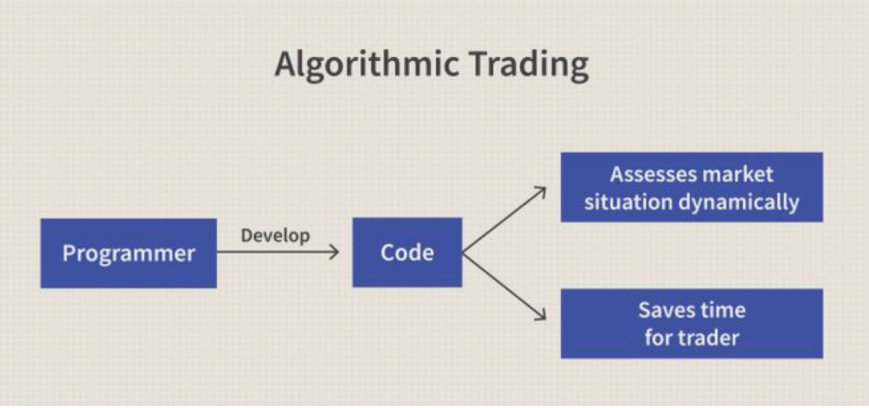
Usage based insurance products in motors and health insurance have been getting attention recently. Pay-As- You-Drive (PAYD), Pay-How-You Drive (PHYD) policies,Pay-As-You-Live (PAYL) policies in health insurance are scalable and helps in studying customer behavior. The penetration of Robo Advisory in the insurance is gaining quite popularity. Many firms are adopting this technology. But I do not favor this, as insurance is the most risky subject with lots of questions in it. A robo may not understand completely as it involves Natural language processing to understand it. People do not like reading long docs as well to understand, instead I would recommend every health insurance company have advisors who guides the customer according to their needs and background. This helps build trust and increase sales.

Big data technology is used on the company's trading and risk data.

5. Algorithmic Trading

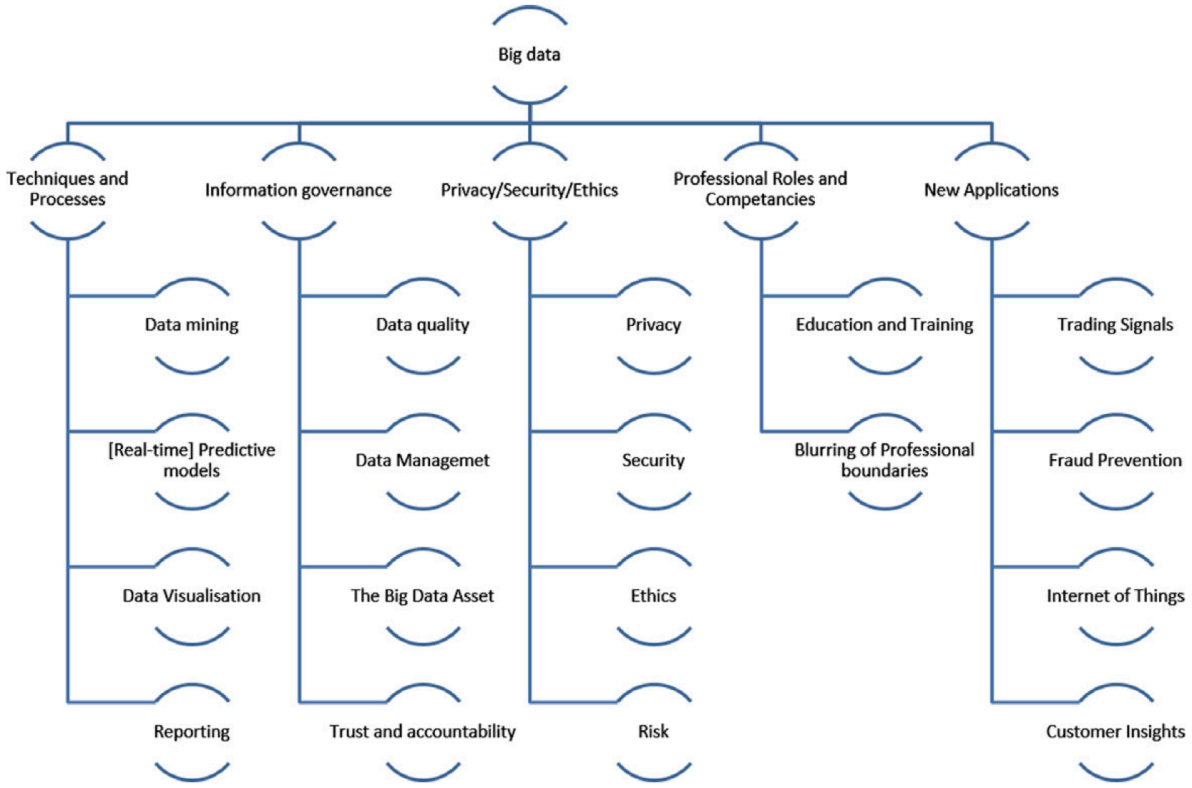
This acts like big data as it composes of multiple hardwares and computers to store and process the data.Many banks used Hadoop for managing unstructured data in 2013. A mix of Hadoop, NoSQL, and predictive analytics of big data has accommodated all its operations areas, and technology is now seen as vital in the act of Deutsche bank decision making and strategic management.

Algorithmic Trading is the automated process that enables computer programs to execute financial trades at speeds and frequencies that a human trader cannot. Unlike decision making, which can be influenced by varying sources of information, human emotion and bias, algorithmic trades are executed solely on financial models and data. Algorithmic trading makes use of complex formulas, combined with mathematical models and human oversight, to make decisions to buy or sell financial securities on an exchange. Stock trader attempts to profit from the purchase and sale of securities such as stock shares with help of algorithmic trading. This algorithmic trading is also beneficial in stocks and out this helps to prevent the time lag in two different areas.



6. Big Data in Accounting

Accounting and finance data is a subset of enterprise data, which includes broader operational and transactional data that can be used for analysis and forecasting.



Identifying meaningful trends and insights from financial and non-financial data, intelligent visualization and presentation of data and using data to improve performance are the three main areas where techniques are applied. The role of IOT in gathering information for the big data is enormous. Every minute, every second huge data are collected, touching all four v's volume, variety, velocity and veracity problems.

The ability of big data to implement real-time predictive modeling makes it excellent choice for handling accounting and financial related matters. i.e ability to make predictive modeling based on historical data.
The reference 12) doesn't explain much in depth with numbers with respect to big data in accounting. It is definitely similar to financial or investment but accounting also deals with taxations, corporate tax, income tax etc. Here's how big data can leverage its capabilities has not been explored in depth.

One of the things that I always thought was to avoid the banks from getting the bad loans. Bad loans are one of the worst performance that indicates from the bank. The NPA which is also called is the non-performing assets indicates the number of bad loans or the bad interest that the people individual or the corporate were unable to pay to the bank which results in the loss circulation off amount with the citizens of this country. Big data's real time predictive modeling can really help the banks to overcome these situations and avoid funding the bad loans to the people/group of corporates based on their brand names. Big data according to the location and the current trend of the market it can store and analyze all the data around the globe and this can update the real time friend of the business. Generally in the banks as I know that charted accountants are responsible for sanctioning of the big loans to the corporate's or an individual, these people are good at mathematics and statistics but they are not the astrologers. Hence predictive modeling with all the Data accumulated from the globe helps to foresee the future in advance and prevent bad loans occurring. This is my individual analysis.

Conclusion

I would like to conclude by saying that big data has an excellent usage on financial sector such as banking stocks accounting entrance sectors etc. there are either a couple of areas where the people need to trust as the data can be very sensitive. The market is very huge for the big data and I see that it would capture a lot of wealth due to its technologies in the future along the both east and west coast of the globe. On-demand cloud computing, huge storage of areas from Hadoop etc. have will have a lot of usage is in the upcoming years and slowly that itself becomes the part of the culture with the date as well. Data engineering and data scientist can never apply their entire skills without the use of big data technologies. The payments the banking sector's the stocks, investments, insurance, the volume of these data collected its velocity it's a variety really requires the usage of big data and this will also see a great

increase in the revenue in the near future. Article Reference 13 shows the future of big data.

References

1) J. Bollen, H. Mao and X. Zeng, Twitter mood predicts the stock market, 2011.

2) . Bhakti G. Deshmukh, Premkumar S. Jain and M. S. Patwardhan, Spin-offs in Indian Stock Market owing to Twitter Sentiments Commodity Prices and Analyst Recommendations, 2016.

3) https://ieeexplore.ieee.org/abstract/document/7840595/references#references

4) https://ieeexplore.ieee.org/abstract/document/8244943

5) https://assets.researchsquare.com/files/rs-573323/v1/00894662-d20d-4f7c-bcb5-86d9b256f987.pdf?c=1631884261

6) https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8012939&casa_token=QKV3d1yl7QwAAAAA:vRiOeFjPGkh6LCRRQegbhUxZPbGklDRuWrpaoADmrZLE1HWN5V2Pmp4kAWv_X3wyMzeiR_LapA&tag=1

7) https://register.eiopa.europa.eu/Publications/EIOPA_BigDataAnalytics_ThematicReview_April2019.pdf

8) https://www.abi.org.uk/news/news-articles/2019/08/detected-insurance-frauds-in-2018/

9) https://onlinelibrary-wiley-com.ezproxy.gl.iit.edu/doi/pdf/10.1002/9781119522225

10) https://onlinelibrary-wiley-com.ezproxy.gl.iit.edu/doi/pdf/10.1002/9781119489368

11) https://web-s-ebscohost-com.ezproxy.gl.iit.edu/ehost/pdfviewer/pdfviewer?vid=1&sid=849cdad9-9b68-4ac9-b288-7df5239bbb7d%40redis

12) https://web-p-ebscohost-com.ezproxy.gl.iit.edu/ehost/pdfviewer/pdfviewer?vid=1&sid=52b31caf-da90-4d3c-9a11-8fb0702d7798%40redis

13) https://www.itransition.com/blog/the-future-of-big-data