Fraud Detection and Analysis for Insurance Claim using Machine Learning

1st Abhijeet Urunkar
Dept. of Computer Science and Engineering
Walchand College of Engineering
Sangli, India
abhijeet.urunkar@walchandsangli.ac.in

2nd Amruta Khot
Dept. of Information Technology
Walchand College of Engineering
Sangli, India
amruta.khot@walchandsangli.ac.in

3rd Rashmi Bhat
Dept. of Computer Engineering
St. John College of Engineering and Management
Palghar, India
rashmib@sjcem.edu.in

4th Nandinee Mudegol
Dept. of Computer Science and Engineering
Walchand College of Engineering
Sangli, India
nandinee.mudegol@walchandsangli.ac.in

***Abstract*—Insurance Company working as commercial enterprise from last few years have been experiencing fraud cases for all type of claims. Amount claimed by fraudulent is significantly huge that may causes serious problems, hence along with government, different organization also working to detect and reduce such activities. Such frauds occurred in all areas of insurance claim with high severity such as insurance claimed towards auto sector is fraud that widely claimed and prominent type, which can be done by fake accident claim. So, we aim to develop a project that work on insurance claim data set to detect fraud and fake claims amount. The project implement machine learning algorithms to build model to label and classify claim. Also, to study comparative study of all machine learning algorithms used for classification using confusion matrix in term soft accuracy, precision, recall etc. For fraudulent transaction validation, machine learning model is built using PySpark Python Library.**

Keywords—Machine Learning Algorithm, PySpark, Fraud Case detection, classifications

I. INTRODUCTION

Insurance fraud is a claim made for getting improper money and not actual amount of money from insurance company or any other underwriter. Motor and insurance area unit two outstanding segments that have seen spurt in fraud. Frauds is classified from a supply or nature purpose of read. Sources is client, negotiator or internal with the latter two being a lot of essential from control framework purpose of reads.

Frauds cowl vary of improper activities that a private might commit so as to attain the favorable outcome from an underwriter. Frauds is classified into nature wise, for example, application, inflation, identity, fabrication, contrived, evoked accidents etc. This could vary from staging incident, misrepresenting matters as well as pertinent members and therefore reason behind finally the extent of injury occurred. Probable things might embrace packing up for a state of affairs that wasn't lined beneath the insurance. Misrepresenting the context of an event. This might embrace transferring blames to the incidents wherever the insured set is accountable, failure to require approved the security measures. Increased impact of the incident .Inflated measure of the loss occurred through the addition of not much related losses or/and attributing inflated price to the increased losses[1][2][3].

II. PROBLEM STSTEMENT

The traditional method for the detecting frauds depends on the event of heuristics around fraud indicators. Supported these, the selection on fraud created is said to occur in either of situations like, in certain things the principles are shown if the case should be interrogated for extra examination. In numerous cases, an inventory would be prepared with scores for various indicators of the occurred fraud. The factors for deciding measures and additionally the thresholds are tested statistically and periodically recalibrated. Associate aggregation and then price of the claim would verify necessity of case to be sent for extra examination. The challenge with above strategies is that they deliberately believe on manual mediation which might end in the next restrictions:

1. Inability to perceive the context-specific relationships between the parameters (geography, client section, insurance sales process) which may not mirror the typical picture.
2. Constrained to control with the restricted set of notable parameters supported the heuristic knowledge – whereas being aware that a number of the opposite attributes might conjointly influence the decisions.
3. Reconstruction of the given model is that the hand operated exercise that need to be conducted sporadically to react dynamic behavior. Also to make sure that the model gives feedback from the examinations. The flexibility to manage this standardization is tougher.
4. Incidence of occurrence of fraud is low - generally but 1percent of claims area unit classified.
5. Consultations with business specialists point out that there is not a typical model to determine the model exactly similar to the context

A. Motivation

Ideally, businesses ought to obtain the responses to prevent fraud from happening or if that is out of the question, to watch it before important damage is finished at intervals

the strategy. In most of the companies, fraud is understood entirely once it happens. Measures are then enforced to forestall it from happening over again. At intervals the given time that they can't resist at different time intervals, but Fraud detection is that the most effective suited issue for removing it from the atmosphere and preventing from continuance once more.

B. *Significance of the Problem*

Knowing a risk is that the beginning in bar, associated intensive assessment offers the lightness that want. This is typically usually performed exploitation varied techniques, like interviews, surveys, focus teams, feedback conducted anonymously, detailed study of record and analysis to spot traffic pumpers, service users, and subscription scam which are different fraudulent case. The association of Certified Fraud Examiners offers a detailed guide to follow. This can be usually alleged to be a preventive methodology, fraud analysis and detection is associate certain consequence of associate intensive risk evaluation. Recognize and classify threats to fraud in knowledge technology and telecommunications sector stereotypically yield the shape of the chances like:

- Records showing associate degree inflated rates in calls at associate degree surreal time of day to associate degree uncertain location or far-famed fraud location.
- Unusual Dialing patterns showing one variety being referred to as additional of times by external numbers than job out.
- Increased calls created in an exceedingly day than the minute's allotted per day, that might indicate an account has been hacked or shared

C. *Major Contribution*

- To compare machine learning algorithms: LR, XGB, DT, RF and SVM.
- To construct a model that predict transactions could be fraudulent with high accuracy.
- To detect if an insurance claim is fraudulent or not.
- To analyze the performance of fraud detection algorithm

III. Literature Review

Machine learning is usually abbreviated as metric capacity unit. The study of machine learning includes computers with the implicit capability to be trained whereas not being expressly programmed. This capacity unit focuses on the expansion of pc programs that has enough capability to alter, that square measure once unprotected to the new information. Metric capacity unit algorithms square measure generally classified into 3 main divisions that square measure supervised learning, unattended learning and reinforcement learning. Data processing a neighborhood of machine learning has advanced considerably within the current years. Data mining focuses at analysing the whole data obtained. Furthermore data processing makes an attempt to seek out the realistic patterns in it. On the contrary, within the different of getting the knowledge for world understanding is within the processing applications like machine learning, it uses the knowledge to locate patterns in information and improvise the program actions thereby. Mainly within the supervised machine learning is that the objective of deducing which means from label on the information used for the coaching. The coaching information consists of a group of coaching samples. Just in case of supervised learning, every instance are often a base which incorporates Associate in Nursing input object that's considered the vector and also the output features a worth that acts as an indicator to run the model. A supervised learning rule initially accomplishes a groundwork task from the sample information then tries to construct a short lived perform, therefore it will plot new input vectors. The supervised learning algorithms square measure conspicuously employed in large choice of application areas. Associate in Nursing best setting altogether the chance assist the rule to accurately mark the class labels for close instances and therefore a similar aspires supervised learning rule to chop back from the knowledge to the enclosed objects in terribly good manner[4][5][6].

The literature review in tabulated form is as follows:

TABLE I. Machine Learning Algorithm Comparision

[7]

Algorithm	Problem Type	Average Prediction Accuracy	Training speed	Prediction Speed	Performance well with small number of object?	Feature Might Need Scaling?
KNN	Either	Lower	Fast	Depends	No	Yes
Logistic Regression	Classification	Lower	Fast	Fast	Yes	No
Support Vector Machine	Either	Lower	Fast(excluding feature extraction)	Fast	Yes	no
Decision Tress	Either	lower	Fast	Fast	No	No

Random Forests	Either	Higher	Fast	Moderate	No	No
XGB	Classification	Higher	Fast	Fast	No	Yes

IV. Proposed Method/Algorithm

The following is the proposed method of the model development:

- Different models are tested on the dataset once it is obtained and cleaned.
- On the basis of the initial model performance, different features of the model are engineered and tested again.
- Once all the options area unit designed, the model is made and run victimisation completely different completely different values and victimisation different iteration procedures.
- A predictive model is created that predicts if an insurance claim is fraudulent or not.
- Binary Classification task takes place which gives answer between YES or NO. This report deals with classification algorithm to detect fraudulent transaction.

A. Proposed System

The influence of the feature engineering, feature choice parameter modification area unit explored with an aim of achieving superior prophetic performance with superior accuracy. The assorted machine learning techniques area unit utilized in the development of accuracy of detection in unbalanced samples. As a system, the info are divided into 3 completely different segments. These area unit loosely coaching, testing and validation.

The algorithmic program is trained on partial set of knowledge and parameters. These area unit later changed on a validation set. This may be studied for evaluation and performance on the particular testing dataset. The high acting models area unit formerly tested with numerous random splits of knowledge. This helps to confirm the consistency in results the approach discussed above comprises of three layers.

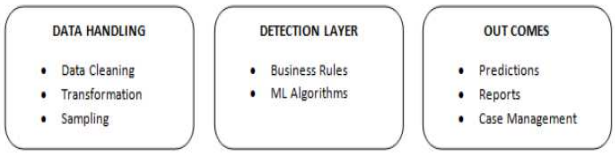


Fig. 1. Layers for Model Building

B. System Architecture

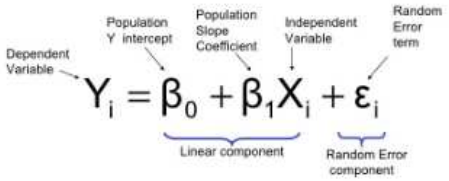
Machine learning model is built with different algorithms that is trained by information and data set provided which predict new classification as "fraud" or "not" These algorithms implemented for building model that is trained using historical data and that predict unseen data with most matching features. And then model is tested and validated to evaluate its performance. After the calculations comparison is made.

For automobile insurance fraud detection supply regression shows the higher accuracy. Logistic regression evaluates the connection among Y "Label" and also the X "Features" by assessing possibilities employing a supply perform. The model predicts a likelihood that is employed to predict the label category. A supply perform or supply curve may be a common curve with equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (1)$$

$$E(Y/X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad (2)$$

Where,



To implement the Logistic Regression using Python, we set the following steps:

- Data Pre-processing step: In this step, the data is ready in order that are often employed in code with efficiency. Extraction of the dependent and freelance variables from the given dataset. Then the dataset is split as coaching and checking victimisation train test split module from sklearn library. Feature scaling is completed therefore on get correct results of predictions
- Fitting Logistic Regression to the Training set: LogisticRegression category of the sklearn library is employed. Classifier object is made and accustomed work the model to the supply regression Predicting the test result: The model is well trained on the training set, the result is predicted by using test set data.
- Test accuracy of the result: Confusion matrix is employed to judge the check accuracy. In this model of fraud detection, the prediction is completed therefore on check if deceitful dealings is claimed as deceitful and the other way around.
- Visualizing the test set result: Adjust the model fitting parameters, and repeat tests. Adjust the model fitting parameters, and repeat tests. Adjust the options or machine learning algorithmic program and repeat tests.

The Methodology of this project is illustrated in below figure:

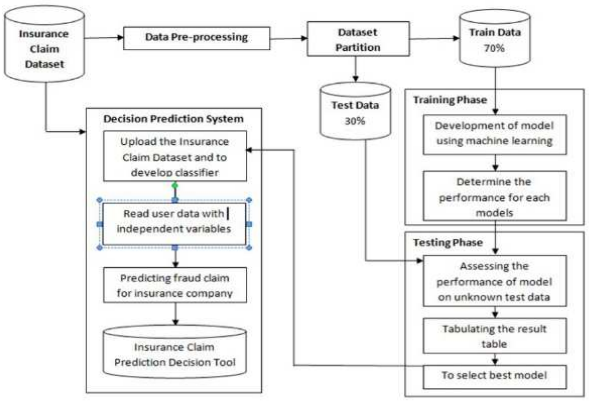


Fig. 2. Prediction Model

C. Implementation Model

Pyspark:

Fraud detection in car Insurance claims are determined employing a python module PySpark MLlib . It's a machine-learning library. it's a wrapper over PySpark Core to try to knowledge analysis exploitation machine-learning algorithms. It works on distributed systems and is ascendable. The implementations of classification, clustering, rectilinear regression, and alternative machine-learning algorithms are found in PySpark MLlib. A repo is employed in PySpark and that they ar ready for giant CSV file process in standalone mode. Particularly, the employment of the 'spark.ml' module was favored because the RDD-based MLLIB library goes to be deprecated.

Scikit-Learn:

Scikit-learn is the most helpful library for machine learning in Python. The sklearn library contains loads of economical tools for machine learning and applied math modeling as well as classification, regression, clump and spatiality eduction.

V. Result and Analysis

Different measures can be used to evaluate and analyse the Model Performance.
Some of the measures used in this project are:

TABLE II. PRECISION ANALYSYS

Model	Recall	Precision	F1 Score
Logistic Regression	79	90	83
XGB	74	89	81
Decision Tree	66	79	71.86
KNN	65	75	68
`Forest Tree	45	77	56

In this analysis, several factors were known which can facilitate to spot for associate degree correct distinction between fraud transactions and non-fraudulent transactions that helps to predict the presence of fraud within the given transactions. Once completely different input datasets are used the Machine Learning models performed at variable performance levels. By considering average F1 score, model rankings are obtained. Higher the F1 score, higher the performance of the model. The analysis indicates that the Adjusted Random Forest formula and changed random below sampling formula provides best performance models.

However, it cannot be assumed that order of prophetical quality would be replicated and might differ for alternative datasets. Once discovered it's complete that within the dataset samples, the models with datasets that are feature made, performs well.

A. Training and Testing Phase

In this analysis, several factors were known which can facilitate to spot for associate degree correct distinction between fraud transactions and non-fraudulent transactions that helps to predict the presence of fraud within the given transactions. Once completely different input datasets are used the Machine Learning models performed at variable performance levels. By considering average F1 score, model rankings are obtained. Higher the F1 score, higher the performance of the model. The analysis indicates that the Adjusted Random Forest formula and changed random below sampling formula provides best performance models.

However, it cannot be assumed that order of prophetical quality would be replicated and might differ for alternative datasets. Once discovered it's complete that within the dataset samples, the models with datasets that are feature made, performs well. Obtained during training and testing phase. Depending on various features the trends are analyzed and hence used to decide the best model among the various Machine Learning classifiers. The following figures gives some of the graphical representation of the results.

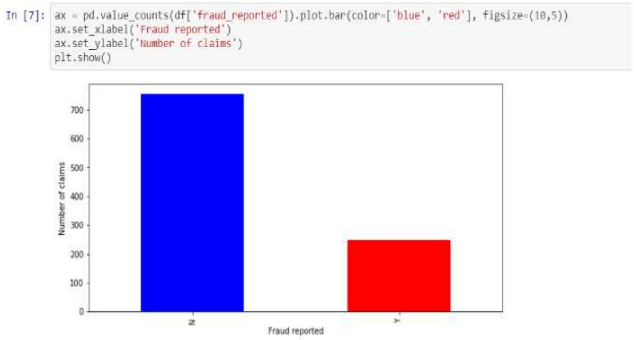


Fig. 3. Fraud Reported

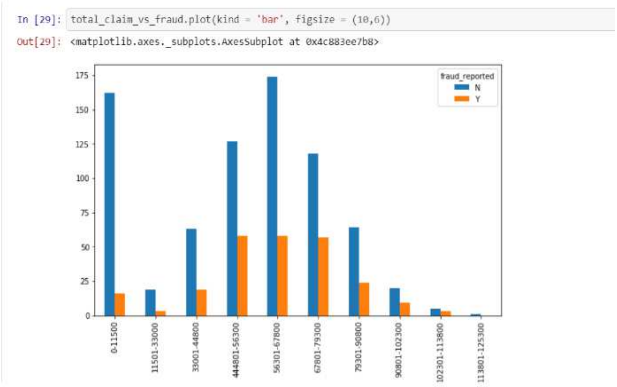


Fig. 4. Total Claim

B. *Feature Selection*

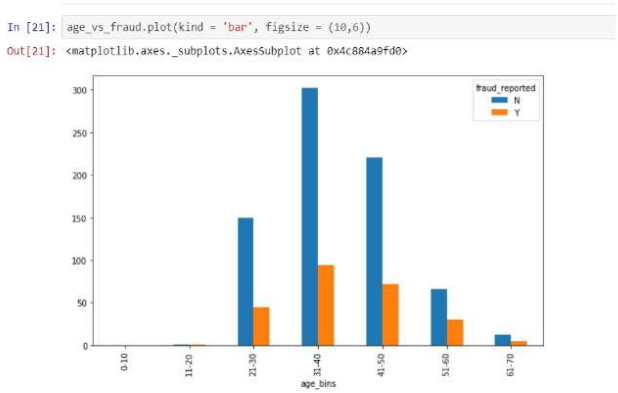


Fig. 5. Age Bins

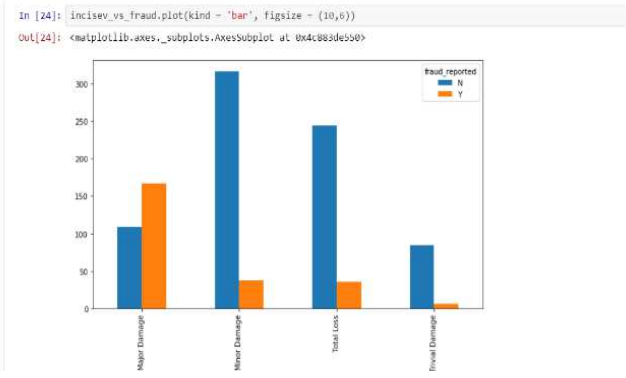
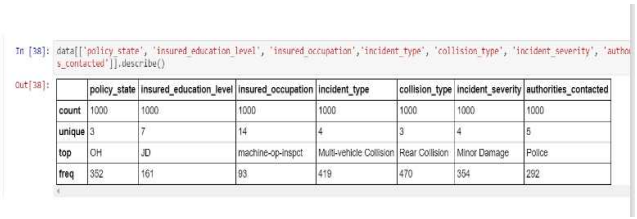


Fig. 6. Incident Sevirty

The following figures gives idea about feature selection. Here we drop column insured education level, insured occupation, authorities contacted since they have very high unique values which will lead to higher number of independent states.



In [38]: data[['policy_state', 'insured_education_level', 'insured_occupation','incident_type', 'collision_type', 'incident_severity', 'authorities_contacted']].describe()

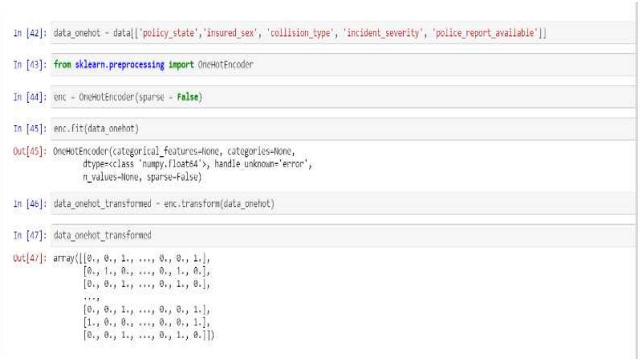
Out[38]:

	policy_state	insured_education_level	insured_occupation	incident_type	collision_type	incident_severity	authorities_contacted
count	1000	1000	1000	1000	1000	1000	1000
unique	3	7	14	4	3	4	5
top	OH	JD	machine-op-inspct	Multi-vehicle Collision	Rear Collision	Minor Damage	Police
freq	352	161	93	419	470	354	292

Fig. 7. Feature Selection

C. *One Hot Encoding:*

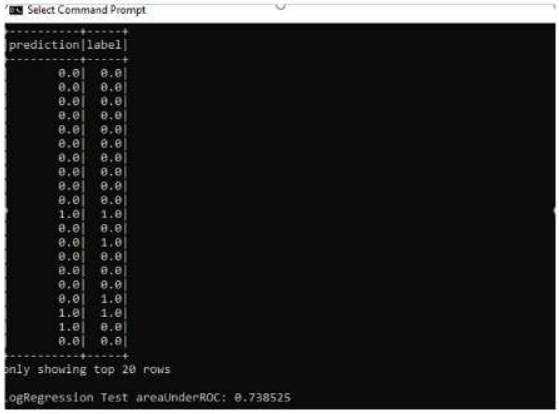
The following figures gives idea about converting all categorical values to numerical using One Hot Encoding. On Hot Encoding allows the representation of categorical data to be more expressive. It is required because many machine learning algorithm unable to work and give expected results with categorical data.



```
In [42]: data_onehot = data[['policy_state','insured_sex', 'collision_type', 'incident_severity', 'police_report_available']]
In [43]: from sklearn.preprocessing import OneHotEncoder
In [44]: enc = OneHotEncoder(sparse = False)
In [45]: enc.fit(data_onehot)
Out[45]: OneHotEncoder(categorical_features=None, categories=None,
               dtype=<class 'numpy.float64'>, handle_unknown='error',
               n_values=None, sparse=False)
In [46]: data_onehot_transformed = enc.transform(data_onehot)
In [47]: data_onehot_transformed
Out[47]: array([[0., 0., 1., ..., 0., 0., 1.],
                [0., 1., 0., ..., 0., 1., 0.],
                [0., 0., 1., ..., 0., 1., 0.],
                ...,
                [0., 0., 1., ..., 0., 0., 1.],
                [1., 0., 0., ..., 0., 0., 1.],
                [0., 0., 1., ..., 0., 1., 0.]])
```

Fig. 8. Data Transformation into Numerical Data

Output of the model:



```
Select Command Prompt
+----------+-----+
|prediction|label|
+----------+-----+
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       1.0|  1.0|
       0.0|  0.0|
       0.0|  1.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  0.0|
       0.0|  1.0|
       1.0|  1.0|
       1.0|  0.0|
       0.0|  0.0|
+----------+-----+
only showing top 20 rows

LogRegression Test areaUnderROC: 0.738525
```

Fig. 9. Ouput for Logistic Regression Model

Input: Auto insurance fraud detection dataset containing 1000 records and 35 features

Output:

- Table comparing the actual results and the predicted results of the model, Where 0.0 stands for 'no fraud' and 1.0 stands for 'fraud'.
- Accuracy of Logistic Regression model.

VI. Conclusion and Future Work

The machine learning models that square measure mentioned which square measure applied on these datasets were able to determine most of the fallacious cases with low false positive rate which suggests with cheap exactness. Certain knowledge sets had severe challenges around data quality, resulting in comparatively poor levels of prediction.

Given inherent characteristics of varied datasets, it would not be sensible to outline optimum algorithmic techniques or use feature engineering process for a lot of higher performance. The models would then be used for specific business context and user priorities. This helps loss management units to specialize in a replacement fraud situations and then guaranteeing that models square measure adapting to spot them. However, it might be cheap to counsel that supported the model performance on back-testing and talent to spot new frauds, the set of models work the cheap suite to use within the space of the insurance claims fraud detection.

VII. References

[1] K. Ulaga Priya and S. Pushpa, "A Survey on Fraud Analytics Using Predictive Model in Insurance Claims," *Int. J. Pure Appl. Math.*, vol. 114, no. 7, pp. 755–767, 2017.

[2] E. B. Belhadji, G. Dionne, and F. Tarkhani, "A Model for the Detection of Insurance Fraud," *Geneva Pap. Risk Insur. Issues Pract.*, vol. 25, no. 4, pp. 517–538, 2000, doi: 10.1111/1468-0440.00080.

[3] "Predictive Analysis for Fraud Detection." https://www.wipro.com/analytics/comparative-analysis-of-machine-learning-techniques-for-%0Adetectin/.

[4] F. C. Li, P. K. Wang, and G. E. Wang, "Comparison of the primitive classifiers with extreme learning machine in credit scoring," *IEEM 2009 - IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2, no. 4, pp. 685–688, 2009, doi: 10.1109/IEEM.2009.5373241.

[5] V. Khadse, P. N. Mahalle, and S. V. Biraris, "An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697476.

[6] S. Ray, "A Quick Review of Machine Learning Algorithms," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.

[7] "https://www.dataschool.io/comparing-supervised-learning-algorithms/.".