HOUSE COST ESTIMATION OF BANGLORE REGION USING FEATURE SELECTION ALGORITHM OF MACHINE LEARNING

SINHA, A.[1*] – RAMISH, M.[2]

[1] *Department of Information Technology, Amity University Jharkhand, Ranchi, India.*

[2] *Department of Electronic and Communication Engineering, Amity University Jharkhand, Ranchi, India.*

**Corresponding author*
e-mail: anuragsinha257[at]gmail.com

(Received 20th March 2021; accepted 29th April 2021)

Abstract. AI assumes a significant part from past years in picture recognition, spam redesign, typical discourse order, item suggestion and clinical determination. Present AI calculation helps us in improving security alarms, guaranteeing public wellbeing and improves clinical upgrades. AI framework likewise gives better client assistance and more secure vehicle frameworks. In the present paper we examine about the forecast of future lodging costs that is produced by AI calculation. For the determination of forecast strategies we look at and investigate different forecast techniques. The housing market is a champion among the most engaged in regards to estimating and continues to change. It is one of the superb fields to apply the thoughts of machine learning on the most proficient method to upgrade and predict the expenses with high exactness. The target of the paper is the expectation of the market estimation of a land property. This investigation uses AI calculations as an exploration technique that creates lodging cost expectation models. We make a lodging cost expectation model in perspective on AI calculation models for instance, XGBoost, rope relapse and neural framework on take a gander at their request exactness execution. We in that point suggest a lodging cost expectation model to help a house seller or a land specialist for better data dependent on the valuation of house. Those assessments display that rope relapse calculation, in perspective on precision, dependably beats substitute models in the execution of lodging cost expectation.
Keywords: *cost estimation, machine learning, support vector machine, logistic regression*

Introduction

Information Mining is extricating information or helpful example from huge data sets. Arrangement is one of the information mining functionalities, utilized for tracking down a model for class characteristic which is a component of other trait esteems (Bogin and Doerner, 2019). Choice Tree is an instrument, which can be utilized for Grouping and Prediction. It has a tree shape structure, where every single inside hub addresses test on an property and the branches out of the hub indicates the test results. 80% of the referred to dataset can be utilized as preparing set furthermore, 20% can be utilized as test informational collection. Each record in the dataset means X and Y esteems, where X is a bunch of characteristic qualities and Y is the class of the record which is the last characteristic in the dataset.

Utilizing the preparation set Choice Tree Classifier model is developed and tried with test information to distinguish the exactness level of the classifier. Choice Tree arrangement as demonstrated utilizes isolate and vanquishes methodology for parting the preparation information into subsets by testing property estimation. This includes quality determination gauges; the trait which is to be tried first is the one which is having high data acquire. Same parting measure is recursively performed on the subsets determined (Bogin and Doerner, 2019). The parting cycle of a subset closes when all

the tuples have a place with similar quality esteem or when no leftover credits or occurrences are left with. Choice Tree arrangement needn't bother with any fundamental area information. It can deal with information of high measurements too. Choice Tree Classifiers have great exactness in grouping. When the Decision Tree is framed, new occurrences can be arranged effectively by following the tree from root to leaf hub. Arrangement through Decision Tree doesn't require a lot of calculation. Choice Trees are able to do taking care of both nonstop and Categorical sort of credits. To keep away from age of useless and undesirable rules in Decision Trees, tree ought not be more profound which results in over fitting. Such a tree with over fitting works more exact with preparing information and less precise with test information.

Relapse is an AI contraption that urges you to make assumptions by taking in - from the current quantifiable data - the associations between your objective boundary and a variety of free boundaries. According to this definition, a house's cost depends upon boundaries, for instance, the quantity of rooms, living locale, region, etc. On the off chance that we apply fake sorting out some way to these boundaries, we can register house valuations in a given land locale. The objective element in this proposed model is the cost of the land property and the autonomous highlights are: no. of rooms, no.of washrooms, cover region, developed territory, the floor, age of the property, postal district, scope and longitude of the property. Other than those of the referenced highlights, which are by and large needed for foreseeing the house costs, we have included two other highlights - air quality and crime percentage. These highlights give an important commitment towards foreseeing property costs since the higher estimations of these highlights will prompt a decrease in house costs.

Literature survey

Land is a flourishing undertaking, and hence forth aexhaustive plan is obliged, which can break down various elements as per the market needs of the country, which is representation in this paper. The exploration directed by Bogin and Doerner (2019) calls attention to that evaluating can be expanded by fifteen percent in mainstream locales, while it showed insignificant effects on the edges of the city. In the examination directed by Xiao-zhu and Ling-wei (2013), it shows that area holds a striking sway on the land costs. Subsequently, there is a requirement for a plan that can investigate such patterns and can have a huge bearing on the house the executives plans and can help a financial backer to bankroll the right way.

The paper "A Quick Review of Machine Learning Calculations," by Ray (2019) features the increases and misfortunes of the regularly utilized calculations of machine adapting in particular Support Vector Machine, Stochastic Angle Descent, K-Nearest Neighbors, Logistic relapse, Naive Bayes Algorithm alongside correlations of these calculations regarding exactness, precision, blunder rate, and so on, it means to give a concise comprehension of regularly utilized calculations of AI to tackle relapse, grouping and order inquiries. Furthermore to this, the exploration was done in "Anticipating the Housing Value Direction utilizing Machine Learning Techniques" by Banerjee and Dutta (2017) edifies the relationship among discrete AI calculations and features that arbitrary woods classifier has high precision also, support vector machine is the preeminent classifier to study which is less inclined to information overfitting.

The work of neural organizations improves the effectiveness of calculations utilized in AI as demonstrated in "House Value Prediction Using Machine Learning and Neural Organizations" by Varma et al. (2018). The result is the mean of a few relapse strategies

like direct relapse and supported relapse to build the precision and reduction the mistake pace of the model to forestall the danger of putting resources into some unacceptable spot. They've utilized Google guides to get solid information from this present reality. The exploration directed on mixture relapse for house value expectation by Lu et al. (2017) advances the investigation of different relapse strategy like Ridge, Rope, and Gradient boosting. Relapses like Lasso what's more, Ridge with a wide number of attributes can demonstrate cases thus this can stay away from information overfitting. Here coupling of different relapses is performed preposterous and it was seen that the Gradient boosting and crossover rope delivered the greatest score. Troupe learning practices different calculations of AI to achieve unrivaled execution than could be accomplished from any of the constituent calculations exclusively.

Ghosalkar and Dhage (2018) use the peril premium learned Likewise those Contrast the center of Different long stretch eagerness rates and the specialists' longings over future brief rates as data variable to anticipating what's to come heading for house costs. They derive that masters Also examiners could utilize enough those larger part of the information given by those venture rate risk premium today so as ought to assess the probability from asserting securing underneath design S&P CS-10 rundown three months ahead. Sampathkumar et al. (2015) likewise estimate the new slump in genuine house value development rates for the twenty biggest U.S. states. The creators utilize Spatial Bayesian VARs (BVARs), in light of on month to month genuine house value development rates, to figure their decline over the time frame 2007:01 to 2008:01. They find that BVAR models are exceptional in guaging the future heading of genuine house costs, however they fundamentally disparage the decrease. They trait this under-expectation of the BVAR models to the absence of any data on basics in the assessment interaction.

Alfiyatin et al. (2017) extend the people same assessment on the 20 generally astounding experienced with metropolitan decay because of de industrialization, headway created organization struggle. States dependent upon ARDL models looking at state, territorial also public level factors. When once more, those creators degree comparative completions on the truth that alliance estimates concerning for independent leeway structure. Au (2018) utilize the risk premium decided likewise the people intricacy those white collar for various entire arrangement energy rates and the specialists' longings In future transient rates as information variable with anticipating what's with go going to house costs. They interpret that aces likewise specialists might use enough the people a lot of the data accommodated by the people financing rate risk premium today with the objective Similarly as should further reinforcing evaluate those likelihood from affirming getting under plan S&P CS-10 summary three months ahead.

Materials and Methods

Data set

The current plan utilizes the information from Kaggle.com, and the dataset has been utilized from the asset open by that web application (*Figure 1*). The dataset picked has 22 credits whereupon the different calculations are tried and prepared. These credits range from various kinds of working to the square feet of the house. It comprises of the utilities accessible suchlike power, water, and Gas. Significant boundaries that enormously influence the cost of a property, for example, the quantity of rooms, generally nature of the house, distance from the primary street, and the region, are

completely analyzed in the dataset positions such as Data Scientist, Software Developer, Web Developer were gathered. At last, the resumes appended to the messages were additionally utilized for testing the model.

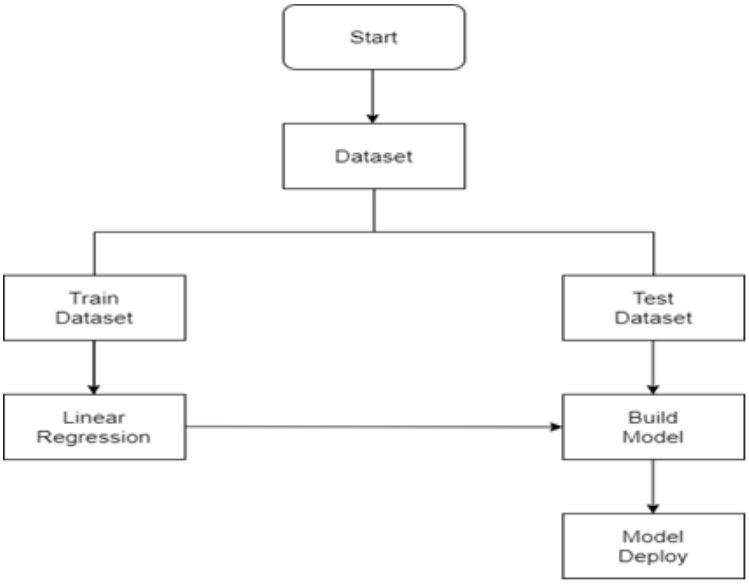

Figure 1. *The proposed method used in this study.*

Algorithm

The regression algorithm of SVM

SVM was proposed in 1995 dependent on factual learning hypothesis (Tan et al., 2009). Contrasted and the customary AI strategies, the AI calculations at present stage are more thorough in rationale and more extraordinary in speculation execution. SVM is developed based on VC measurement hypothesis and construction hazard least standard, seeking after the best equilibrium point between the learn capacity and model intricacy. In the issues of little example, nonlinearity and high measurement, SVM has critical benefits. Little example alludes not to without a doubt the quantity of tests, but rather the quantity of tests which SVM calculation requires is generally little to the intricacy of the issue. Taking care of the nonlinear issues is the center of SVM technique. By the presentation of portion capacity and slack variable SVM calculation cunningly tackled the issue of direct indivisible. Piece work is a capacity that fulfills the Mercer condition. The intricacy of the computation is successfully decreased by the piece work. The motivation behind why SVM has extremely huge benefits in tackling high measurement issues is that SVM doesn't have to utilize the entirety of the examples in managing issues.

SVM is for the most part used to take care of the issues of order of the examples of various classes and the relapse of the examples. The grouping issue principally alludes to looking for a hyperplane in the higher dimensional space to isolate out the examples of various classes. For SVM, the numerous order can be settled through developing two classifiers. SVM relapse is to foresee the solid worth based on various example attributes (Banerjee and Dutta, 2017; Jiang and Li, 2010). The preparation informational collection is characterized as, where is the info information and is the relating yield information. The relapse capacity can be communicated as follows (Eq. 1):

$$\min_{z^*,b,\xi^{(*)}} \quad \frac{1}{2}\left\|z^*\right\|^2 + F\sum_{i=1}^{i}\left(s_i + s_i^*\right)$$

$$\text{s.t.} \quad \begin{cases} \left(\left(z^* \cdot x_i\right) + l\right) - y_i \le \varepsilon + s_i, & i = 1,\ldots,q \\ y_i - \left(\left(z^* \cdot x_i\right) + l\right) \le \varepsilon + s_i^*, & i = 1,\ldots,q \\ s_i^* \ge 0, & i = 1,\ldots,q, \end{cases} \qquad \text{(Eq. 1)}$$

Where; addresses weight vector and is an ordinary. Both of these boundaries can be gotten from the capacity.

Logistic regression

This methodology plans to address the probability that an event appears dependent on appraisals of test factors that could be either mathematical or straight out. It endeavors to decide the effect of a bunch of factors on a double reaction variable and figures the likelihood of an occasion happening for a haphazardly chose perception against the probability it doesn't happen. It gatherings discoveries by ascertaining the probably hood of an event in a specific class, for instance, ordering whether a site is fake or not.

Sigmoid function

In Equation 2, where x is the information esteem, Y is the yield also, e is Euler's number. The bend of the sigmoid capacity takes after as a S-shaped bend. The main role of utilizing a sigmoid work is that it ranges inside nothing and one. It's especially useful for models where likelihood is to be anticipated as a result and since the likelihood of everything ranges somewhere in the range of nothing and one, it’s the right decision.

$$Y=1/1+e^{(-x)} \qquad \text{(Eq. 2)}$$

Support vector machine

Here in this technique, we complete the arrangement by characterizing the hyperplane among the qualities. Each information thing is deciphered as a point, and they are plot versus n-dimensional space with the end goal that the estimation of each element being the estimation of a specific facilitates. Here the utilization of SVM is to deteriorate the blunder rate and misclassification by perceiving hyperplane with high edge from the information point. SVM are instrumental if there should arise an occurrence of high dimensional space. Singular credits have facilitates which are officially fathomed as help vector. To achieve the information change from the lower-dimensional information space towards higher dimensional information space, we use bit work, which urges to handle such intricate change.

Naïve Bayes

It's a measurable arrangement method used to address issues concerning arrangement. It is a quick, exact, solid calculation and has high precision and speed on huge datasets. Innocent Bayes infers that the presence of a explicit trademark is free of

other qualities being available. For example, if the organic product is orange in shading, ten centimeters in breadth and round in shape, the natural product could be seen as orange. All of these properties, both independently and self-rulingly, lead to the probability that the natural product is orange while paying pretty much nothing respect to whether they depend upon each other, and that is the explanation it is called ' Naive.' Bayes Theorem: The hypothesis gets the probability of an occasion occurring given that another occasion has effectively occurred. Its communicated by the accompanying recipe; P(L|M) – the probability of occasion L happening, given occasion M has happened; P(M|L) – the probability of occasion M happening, given occasion L has happened.

Stochastic gradient descent

It's a generally utilized calculation of AI and structures the premise of neural organizations. The overall idea of slope plummet in AI comprises of consecutively changing boundaries to diminish the expense work. There is a term called group in Gradient Descent, which indicates the complete amount of tests from a dataset used to gauge the angle for every emphasis. Normally, this group is taken to be the entire dataset. The issue emerges when the dataset is immense. In the event that we utilize a normal angle drop enhancement strategy, we should utilize all the tests to finish one emphasis while the Gradient Drop is performed, and it should be rehashed for each cycle until the base is reached. Accordingly, it turns out to be computationally costly, while Stochastic Slope Descent at each stage chooses one case from the dataset at irregular rather than the entire dataset and refreshes angle dependent on that solitary record. The advantage of it is that it's computationally more affordable, and in many situations, it's liked over other slope plunge strategies for streamlining.

K-nearest neighbors

It is a basic, adaptable and clear to carry out directed learning calculation. It deals with the philosophy that comparative perceptions are near one another. It catches the idea of closeness by computing the partition inside two focuses on a diagram. The 'k' in the calculation is a mathematical worth that advises the number of information focuses to consider for taking a vote. To group another point, we surround the point with K number of datapoints and appoint it to the bunch with the greatest number of focuses inside the circle. The ideal method to distinguish the estimation of K is by evaluating a barely any estimations of k prior to choosing one, which lessens the mistake and simultaneously keeps up the precision of the forecast. Low qualities can be loud and dependent upon exceptions. Huge estimations of K smooth over thing however K ought not be so enormous that different classifications will consistently outvote a class with a couple of models.

Results and Discussion

Data cleaning

Data Cleaning is a procedure used to change over non-analyzed information into a dependable informational collection (*Table 1*). In various words, unformatted information gathered from various sources isn't good for examination. In this part, the dataset is stacked, checked for information replication, and repetition, after which the

dataset is cleaned and managed to reduce information irregularity and upgrade the examination. The dataset may fuse missing qualities and invalid qualities, which can cause irregularity and wrong outcomes. It is an essential endeavor to be executed to achieve results with better exactness. The essential goal of cleaning the information is to recognize and dispose of any mistakes to upgrade the information for choice making and examination. Since the calculations can handle as it were mathematical worth, the all out information inside the dataset is named utilizing mark encoder. The dataset amassed is ordered into two divisions: a Training set and Test set for foreseeing the house costs (*Table 2* and *Table 3*). The model developed utilizing various methods of AI is put to the preparing set. The Test set checks the precision of the model. To additional the viability of the model, we have utilized cross-approval.

***Table 1**. Arranging data and feature selection.*

No	Area_type	Availability	Location	Size	Society	Total_ sqft	Bathroom	Balcony	Price
0	Super built-up area	19 Dec 2021	Electronic city phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot area	Ready to move	Chikka Tirupathi	4 bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up area	Ready to move	Uttarahalli	3 BHK	Na	1440	2.0	3.0	62.00
3	Super built-up area	Ready to move	Lingadheeranahalli	3 BHK	Na	1521	3.0	1.0	95.00
4	Super built-up area	Ready to move	Kothanur	2 BHK	Na	1200	2.0	1.0	51.00

***Table 2**. Data used and analyzed of Bangalore region.*

No	Location	Size/BHK	Total_sqft	Bath	Price	Price_per_sqft
0	Electronic city phase II	2	1056.0	2.0	39.07	3699.81
1	Chikka Tirupathi	4	2600.0	5.0	120.00	4615.38
2	Uttarahalli	3	1440.0	2.0	62.00	4305.56
3	Lingadheeranahalli	3	1521.0	3.0	95.00	6245.89
4	Kothanur	2	1200.0	2.0	51.00	4250.00
5	Whitefield	2	1170.0	2.0	38.00	3247.86
6	Old airport road	4	2732.0	4.0	204.00	7467.06
7	Rajaji Nagar	4	3300.0	4.0	600.00	18181.82
8	Marathahallo	3	1310.0	3.0	63.25	4828.24
9	Others	6	1020.0	6.0	370.00	36274.51

***Table 3**. The data output.*

No	Total_sqft	Bath	Price	BHK	The block of Jayanagar
0	2850.0	4.0	428.0	4	1
1	1630.0	3.0	194.0	3	1
2	1875.0	2.0	235.0	3	1
3	1200.0	2.0	130.0	3	1
4	1235.0	2.0	148.0	2	1

Execution measurement

In this model, we use disarray network to decide the exactness, particularity, affectability, f1 factor and accuracy. The disarray grid is a presentation assurance table offered for the order model on a bunch of information. Through chart and different measures, we can effectively grasp the presentation of the information. Each class reports the quantity of wrong and right forecast checks. Two essential sorts: Positive and Negative, have been portrayed and subsequently different results are being portrayed. This include; (1) Positive (P): Observation are positive; (2) Negative (N): Observation are negative; (3) Genuine Positive (TP): Observation are positive, and are

anticipated to be positive; (4) Bogus Negative (FN): Observation are positive, however are anticipated negative; (5) Genuine Negative (TN): Observation are negative, and are anticipated to be negative; and (6) Bogus Positive (FP): Observation are negative, yet are anticipated positive. In equation, (I) Accuracy: We figure precision as Precision = (TP +TN)*100/(TP + TN + FP + FN); (II) Precision: We figure exactness as Exactness = (TP*100)/(TP + FP); (III) Recall: We characterize review as Review = (TP)/(TP +FN); (IV) Sensitivity: We characterize affectability as Affectability = (TP*100)/(TP + FN); and (V) Specificity: We characterize particularity as Affectability = (TP * 100)/(TP + FN).

To assemble stacking strategy to we need to characterize two things: a meta classifier that we need it to gathering with also, the frail students L (*Table 4*). Here the powerless Learners utilized in this proposed structure are Logistic Regression, Extra Tree Stochastic Gradient Descent, Naïve Bayes, SVM and K-NN (*Table 5*). The meta classifier that we use is the democratic classifier which performs delicate democratic to get wanted result in this system.

***Table 4**. The characteristic of stacking strategy*.

No	Model	Best_score	Best_parameters
0	Linear_regression	0.818354	{‘normalize’: False}
1	Lasso	0.687503	{‘alpha’: 1, ‘selection’: ‘random’}
2	Decision_tree	0.719227	{‘criterion’: ‘mse’, ‘splitter’: ‘best’}

***Table 5**. The result of powerless learners*.

Algorithm	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
Logistic regression	95.89	96.64	84.69	96.43
Support vector machine	81.48	81.28	71.40	74.60
Naïve Bayes	95.84	94.80	94.89	96.05
Stochastic gradient descent	81.53	92.37	64.28	98.37
Extra tree	96.78	92.51	89.79	98.37
K-NN	79.12	81.74	5.86	95.76
Voting classifier	98.27	97.98	91.07	98.49

Conclusion

This article utilizes the most key machine learning calculations like choice tree classifier, choice tree relapse and different direct relapse. Work is executed utilizing Scikit-Learn AI device. This work assists the clients with foreseeing the accessibility of houses in the city and furthermore to anticipate the costs of the houses. Two calculations like choice tree relapse and various direct relapse were utilized in anticipating the costs of the houses. Relatively the exhibition of various direct relapse is discovered to be superior to the choice tree relapse in anticipating the house costs. In future the dataset can be set up with more highlights and progressed AI strategies can be for building the house value forecast model.

In this paper, the Decision tree AI calculation is utilized to develop an expectation model to anticipate potential selling costs for any land property. Extra highlights like air quality and wrongdoing rate were remembered for the dataset to help anticipate the costs far and away superior. These highlights are not generally remembered for the datasets of other expectation frameworks, which makes this framework unique. These highlights impact individuals' choice while buying a property, so why exclude it in anticipating house costs. The prepared model is coordinated with the User Interface

utilizing the Flask Structure. The framework gives 89% exactness while anticipating the costs at the land costs.

Acknowledgement

This research study is self-funded.

Conflict of interest

The author confirm there are no conflict of interest involve with any parties in this research study.

REFERENCES

[1] Alfiyatin, A.N., Febrita, R.E., Taufiq, H., Mahmudy, W.F. (2017): Modeling house price prediction using regression analysis and particle swarm optimization. – International Journal of Advanced Computer Science and Applications 8(10): 323-326.

[2] Au, T.C. (2018): Random forests, decision trees, and categorical predictors: The "Absent Levels" problem. – Journal of Machine Learning Research 19: 30p.

[3] Banerjee, D., Dutta, S. (2017): Predicting the housing price direction using machine learning techniques. – In IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) 3p.

[4] Bogin, A.N., Doerner, W.M. (2019): Property renovations and their impact on house price index construction. – Journal of Real Estate Research 41(2): 249-284.

[5] Ghosalkar, N.N., Dhage, S.N. (2018): Real estate value prediction using linear regression. – In Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 5p.

[6] Jiang, L., Li, C. (2010): An empirical study on attribute selection measures in decision tree learning. – Journal of Computational Information Systems 6(1): 105-112.

[7] Lu, S., Li, Z., Qin, Z., Yang, X., Goh, R.S.M. (2017): A hybrid regression technique for house prices prediction. – In IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 5p.

[8] Pow, N., Janulewicz, E., Liu, L. (2014): Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal. – Course project, COMP-598, Fall/2014, McGill University 8p.

[9] Ray, S. (2019): A quick review of machine learning algorithms. – In International conference on machine learning, big data, cloud and parallel computing (COMITCon) 4p.

[10] Sampathkumar, V., Santhi, M.H., Vanjinathan, J. (2015): Forecasting the land price using statistical and neural network software. – Procedia Computer Science 57: 112-121.

[11] Tan, K.C., Teoh, E.J., Yu, Q., Goh, K.C. (2009): A hybrid evolutionary algorithm for attribute selection in data mining. – Expert Systems with Applications 36(4): 8616-8630.

[12] Varma, A., Sarma, A., Doshi, S., Nair, R. (2018): House price prediction using machine learning and neural networks. – In Second International Conference on Inventive Communication and Computational Technologies (ICICCT) 4p.

[13] Xiao-zhu, D., Ling-wei, K. (2013): The land prices and housing prices—Empirical research based on panel data of 11 provinces and municipalities in Eastern China. – In International Conference on Management Science and Engineering 20th Annual Conference Proceedings 6p.