

A
Project Report
On

**Health Insurance Cost Prediction using
IBM Auto AI Service**

Submitted by
SUDHANSHU MATHUR
JAGANNATH UNIVERSITY, JAIPUR

1. INTRODUCTION

1.1 Overview

The main foundational block of health insurance industry is to estimate the future events and measure the associated risk/value of these events, hence it is needless to say that predictive analytics is used widely to determine the risk, insurance premium and enrich overall customer experience. The health insurance industry has always been a slow-moving industry when it comes to adopting the data analytics practices into its business models. With the advent of advanced data analytics technologies, it has become important more than ever to take advantage of such sophisticated analytics to accurately assess and predict the insurance premiums for the insured. Thus, one of the important tasks for health insurance companies is to determine the policy premiums. By using predictive modeling, the insurers can determine the policy premium for the insured based on their behaviors which are indicated by attributes such as age, BMI (Body Mass Index), smoking habits, number of children etc.

1.2 Purpose

Judicious use of predictive analysis has empowered health insurers to improve their premium pricing accuracy, create customized health insurance plans and services, and build stronger customer relationships. Thus, the main goal of this project is to predict the insurance premiums based on the behavioral data collected from the individuals so that insurance companies can make useful and accurate predictions.

The determination of premiums based on the data collected for an individual helps insurance companies in enhanced pricing, underwriting and risk selection. Additionally, it helps in making better decisions, understanding customer needs and be fair to the customers. Acquiring a comprehensive understanding of customer behaviors and habits from historical data helps insurers to anticipate future behaviors and provide the right insurance product and policy premium.

Based on these predictions, they can then evaluate the following decisions and make better judgement calls:

- Which individuals deserve which kind of insurance plan?

- Based upon an individual's behavior, predicting their premium helps in better risk management.

2. LITERATURE SURVEY

2.1 Existing problem

Rich and well educated households typically have both better health (Asfaw, 2003) and better health insurance coverage (Jütting, 2004; Cameron and Trivedi, 1991), but the positive correlation between health and insurance status tells us nothing about the impact of insurance. On the other hand, those in poor health may be more likely to pay for health insurance (Cutler and Reber, 1998; Ellis, 1989), but finding that the insured tend to be sicker would not imply that insurance causes illness. Some studies in wealthier nations find evidence that people with higher expected medical expenditures (measured in a variety of ways across studies) are more likely to buy insurance or pay for health insurance at higher premiums than those with lower expected medical expenditures (e.g. Cutler and Zeckhaus, 1998). However, the extent of adverse selection in health and other insurance is often found to be minimal (e.g. Wolfe and Goddeeris, 1991; Finkelstein and Poterba, 2004) or non-existent (e.g. Finkelstein and McGarry, 2006; Cardon and Hendel, 2001; Cawley and Philipson (1999). There is also some recent evidence of positive selection into health insurance (e.g. Fang et al., 2008).

Relation of health insurance purchase decision and health expenditure is based on the premise that families which have higher chances of requiring hospitalization will have higher probability of buying health insurance. Some other socio economic factors like age, education etc. have also been found to be important factors affecting health insurance purchase In India knowledge and awareness about health insurance could be important factor for health insurance purchase decision. Very few studies have tried to analyse reasons for low penetration of health insurance in India (Wadhawan 1987, Ellis 2000, Bhat and Mavalankar 2001). Some studies have tried to analyse community based health insurance in India. (Devadasan, Ranson et al. 2004, Ahuja 2005. Rao (2004) discusses the issues and challenges for health insurance sector in India. These and other studies have tried to analyse health insurance sector in India, but not much systematic empirical work has been done and this area is largely unexplored. The theory of risk has been

applied extensively to the literature related to health insurance decision (Arrow 1963; Feldstein 1973). Under conditions of consumer rationality and risk averseness, the decision to purchase insurance is made on the basis of expected utility gain.

2.2 Proposed solution

Judicious use of predictive analysis has empowered health insurers to improve their premium pricing accuracy, create customized health insurance plans and services, and build stronger customer relationships. By using predictive modeling, the insurers can determine the policy premium for the insured based on their behaviors which are indicated by attributes such as age, BMI (Body Mass Index), smoking habits, number of children etc.

3. THEORITICAL ANALYSIS

3.1 Block diagram

Understanding Dataset

The dataset is being taken from popular website called Kaggle. This dataset contains the information on individual attributes such as sex, age, smoking habits etc. It has 1338 rows and 7 columns.

Description of columns:

Age – age of primary beneficiary

Sex – gender of the beneficiary. It has two categories: Male or Female

BMI – Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9

Children – Number of children covered by the health insurance / Number of dependents.

Smoker – describing whether a person is a smoker or a non-smoker. It has 2 values: Yes or No

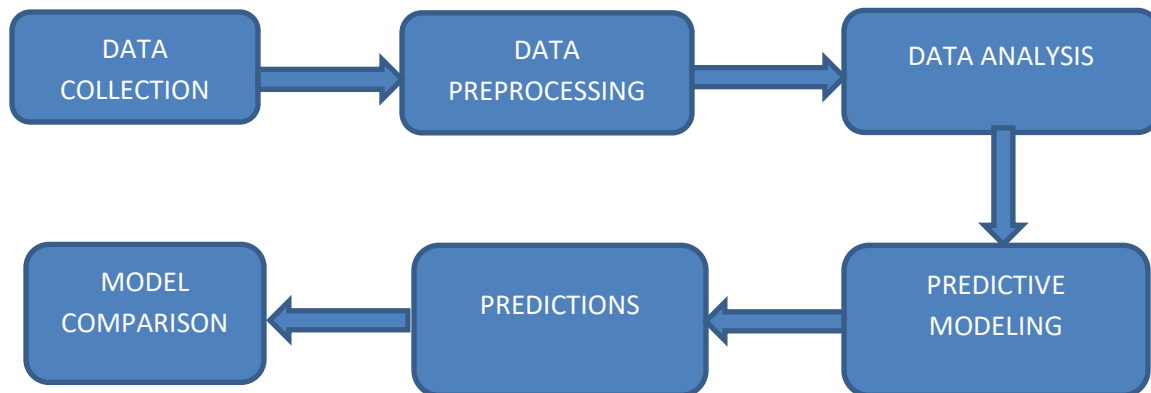
Region – the beneficiary's residential area in the US. It has 4 region values: Northeast, Southeast, Southwest, Northwest

Expenses – Individual insurance premiums billed by health insurance.

The main goal of the project is to predict the insurance premium expenses based upon other

attributes. Therefore, the *dependent variable* is *expense*. All the other attributes are independent variables.

The methodological approach adopted here can be explained by following block diagram:



3.2 Hardware / Software designing

Here, IBM cloud services are used for above mentioned work flow. In particular, IBM Watson Studio is employed here which is a strong tool to prepare data and build models at scale across any cloud. With its open, flexible multi cloud architecture, Watson Studio provides capabilities that empower businesses to simplify enterprise data science and AI.

After selecting an empty project on Watson Studio, we have to associate cloud storage object with the project. It is necessary to use IBM Watson Machine Learning instance with the project. This instance will automatically understand the data and do preprocessing task itself. It also predict the nature of the problem just by getting the target variable. If the target variable is continuous type, it will automatically predict it as regression type problem. It also check for the evaluation metrics like root mean square value (RMSE) etc. The following cloud services are used here:

1. IBM Watson Studio
2. Cloud Object Storage

3. IBM Watson Machine Learning

4. Node Red

EXPERIMENTAL INVESTIGATIONS

After selecting 'expense' as a target variable, IBM Auto AI tool understand it as a regression type Problem. This tool will then identify the suitable algorithms for this problem. Also, it is looking for Evaluation metrics for each algorithm. Various machine learning pipelines are identified by the tool itself. Figure 1 shows the machine learning pipelines for various algorithms and model.

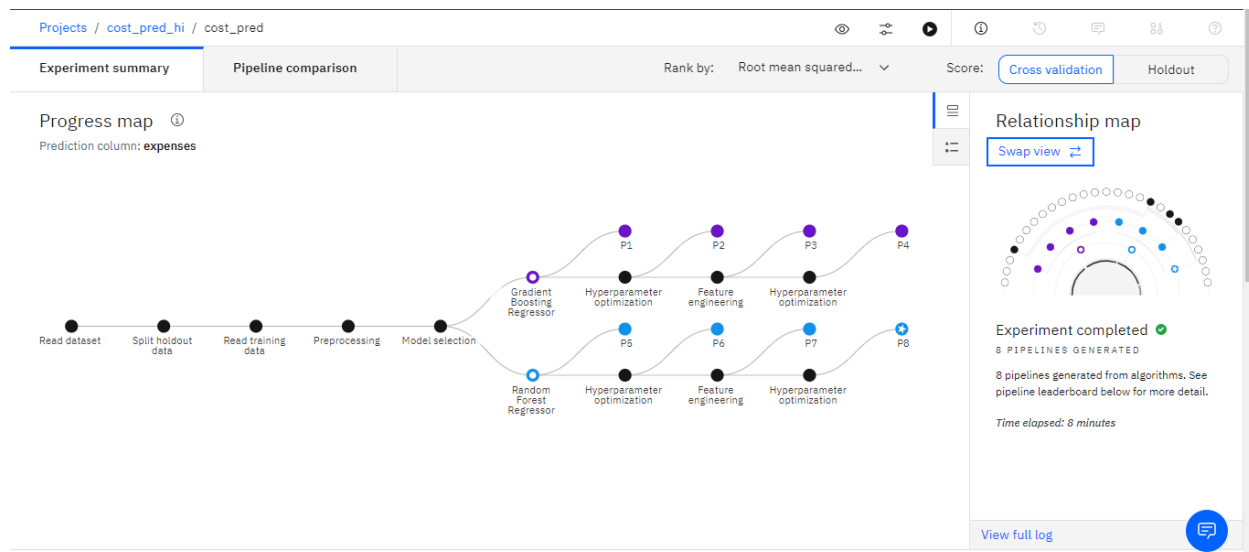


Fig. 1 Machine learning pipelines

There are various evaluation metrics parameters like mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) etc. These parameters are evaluated to check the performance of the model. The flow of each pipeline with the evaluation metrics is shown in the figure given below:

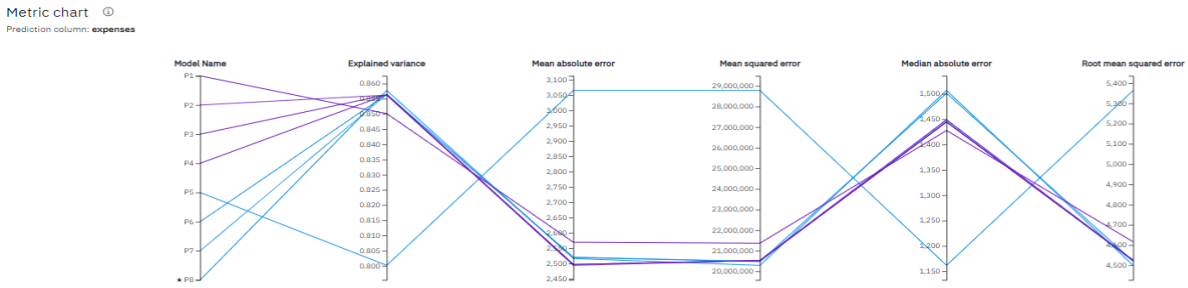


Fig. 2 Metric Chart for different pipelines

RESULTS

It has been observed that there are eight pipelines formed by the tool. Out of which, pipeline number eight performs well over the other models. This is because here the value of RMSE is found minimum among all other pipelines. It comes out to be 4499.864. Here, hyperparameter tuning is also employed and results seems good.

Projects / cost_pred_hi / cost_pred

Experiment summary		Pipeline comparison		Rank by: Root mean squared...	Score: Cross validation	Holdout
Rank	↑	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
★ 1		Pipeline 8	Random Forest Regressor	4499.864	HPO-1 FE HPO-2	00:01:04
2		Pipeline 6	Random Forest Regressor	4520.222	HPO-1	00:00:15
3		Pipeline 7	Random Forest Regressor	4520.483	HPO-1 FE	00:01:38
4		Pipeline 3	Gradient Boosting Regressor	4524.779	HPO-1 FE	00:01:26
5		Pipeline 4	Gradient Boosting Regressor	4524.779	HPO-1 FE HPO-2	00:00:48
6		Pipeline 2	Gradient Boosting Regressor	4525.229	HPO-1	00:00:17
7		Pipeline 1	Gradient Boosting Regressor	4616.347	None	00:00:01
8		Pipeline 5	Random Forest Regressor	5363.872	None	00:00:01

Fig. 3 Comparison of various pipelines

We also obtain the evaluation measures for the selected model. It is shown in figure 4 below where holdout scores and cross validation scores are available.

Model Evaluation Measures ⓘ

TARGET : EXPENSES

	Holdout Score	Cross Validation Score
Root Mean Squared Error (RMSE)	4,483.446	4,499.864
R ²	0.872	0.858
Explained Variance	0.872	0.858
Mean Squared Error (MSE)	20,101,287.613	20,291,289.889
Mean Squared Log Error (MSLE)	0.163	0.180
Mean Absolute Error (MAE)	2,430.645	2,516.402
Median Absolute Error (MedAE)	1,341.359	1,506.721
Root Mean Squared Log Error (RMSLE)	0.404	0.423

Fig. 4 Model evaluation measures

By the designed predictive model, it is observed that smoking habit makes a greater impact on the insurance premium calculation. The BMI is also responsible for the higher expenses of health insurance.

Feature Importance ⓘ

TARGET : EXPENSES

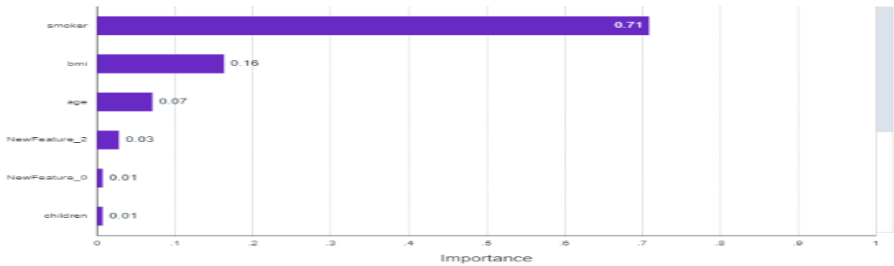


Fig. 5 Feature Importance

Finally, the user interface was created using the cloud foundry application called as Node Red App. The UI can be looks like

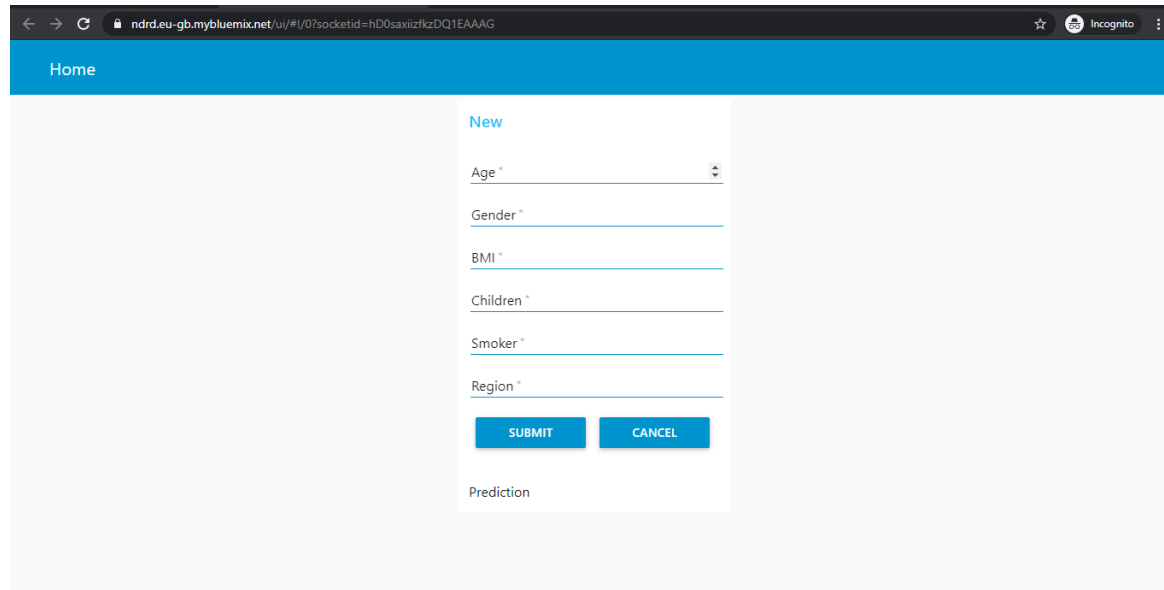


Fig. 6 UI designed for the model

ADVANTAGES, LIMITATIONS AND APPLICATIONS

Using this model, the health insurance providers can then evaluate the following decisions and make better judgement calls like:

- Which individuals deserve which kind of insurance plan?
- How much the premium should be charged based on an individual's behaviors?
- Based upon an individual's behavior, predicting their premium helps in better risk management.
- It helps forge trust between the customer and the insurance company.

Thus, it is important for a health insurance company to collect and analyze the data such as a person's age, BMI, health data to accurately predict the risk and charge accurate premiums to cover that risk.

However, the data did not include any information on an individual's medical costs, the real-time data i.e. data collected from the sensors in the wearable health devices such as fit bits etca. If we take all these types of different data sources into account then we can have a better picture of an

individual's behavior and can more accurately predict the insurance premium charge and the associated risk.

CONCLUSION

Based on the findings of the project, it can be said that predictive modeling has tremendous benefits for the health insurance industry in determining how much the premium should be charged to the insured based upon his/her behaviors and health habits. Health insurance companies can then accurately charge the premium based upon a specific individual's attributes. This will not only help the individuals in getting charged the right amount of premium for their health insurance but will also help in forging better relationships and a level of trust between the insurance company and the insured.

FUTURE SCOPE

However, there are certain limitations which are the scope of further studies. The data did not include any information on an individual's medical costs, the real-time data i.e. data collected from the sensors in the wearable health devices such as fit bits etcetera. If we take all these types of different data sources into account then we can have a better picture of an individual's behavior and can more accurately predict the insurance premium charge and the associated risk.

BIBLIOGRAPHY

1. Nyce, Charles (2007), Predictive Analytics White Paper(PDF), American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America
2. Conz, Nathan (September 2, 2008), "Insurers Shift to Customer-focused Predictive Analytics Technologies", Insurance & Technology
3. Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley Series in Probability and Statistics, 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679.
4. "Linear Regression (Machine Learning)" (PDF). University of Pittsburgh.
5. Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
6. Liaw A (16 October 2012). "Documentation for R package randomForest" (PDF). Retrieved 15

March 2013.

7. "Artificial Neural Networks as Models of Neural Information Processing | Frontiers Research Topic". Retrieved 2018-02-20.

8. Hoskins, J.C.; Himmelblau, D.M. (1992). "Process control via artificial neural networks and reinforcement learning". *Computers & Chemical Engineering*. 16 (4): 241– 251. doi:10.1016/0098-1354(92)80045-B.

9. Applied Predictive Modelling by Max Kuhn, Kjell Johnson publishes by Springer 2016.

APPENDIX

A. Source code

```
[{"id":"21197ab8.4d52c6","type":"tab","label":"HI Cost prediction project","disabled":false,"info":""},{ "id":"9ed6edcb.9041b","type":"ui_form","z":"21197ab8.4d52c6","name":"","label":"","group":"8e58286.aca2fd8","order":1,"width":0,"height":0,"options":[{"label":"Age","value":"ag","type":"number","required":true,"rows":null},{ "label":"Gender","value":"sx","type":"text","required":true,"rows":null},{ "label":"BMI","value":"bmi","type":"number","required":true,"rows":null},{ "label":"Children","value":"ch","type":"number","required":true,"rows":null},{ "label":"Smoker","value":"sm","type":"text","required":true,"rows":null},{ "label":"Region","value":"rg","type":"text","required":true,"rows":null}], "formValue":{"ag":"","sx":"","bmi":"","ch":"","sm":"","rg":""}, "payload":"","submit":"submit","cancel":"cancel","topic":"","x":70,"y":220,"wires":[["885f9cca.d352d","a49c4564.85e1b8"]]}, {"id":"885f9cca.d352d","type":"debug","z":"21197ab8.4d52c6","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"false","statusVal":"","statusType":"auto","x":250,"y":300,"wires":[]}, {"id":"a49c4564.85e1b8","type":"function","z":"21197ab8.4d52c6","name":"pre-token","func":"global.set(\\\"ag\\\",msg.payload.ag)\\nglobal.set(\\\"sx\\\",msg.payload.sx)\\nglobal.set(\\\"bmi\\\",msg.payload.bmi)\\nglobal.set(\\\"ch\\\",msg.payload.ch)\\nglobal.set(\\\"sm\\\",msg.payload.sm)\\nglobal.set(\\\"rg\\\",msg.payload.rg)\\nvar apikey=\\\"h1e93on1hcjm6BFWovVXd-NqbHDI2tFEUTiP8-aQ7zpT\\\";\\nmsg.headers=\\{\\\"content-type\\\":\\\"application/x-www-form-urlencoded\\\"\\}\\nmsg.payload=\\{\\\"grant_type\\\":\\\"urn:ibm:params:oauth:grant-type:apikey\\\",\\\"apikey\\\":apikey\\}\\nreturn msg;\\n\", \"outputs\":1,\"noerr\":0,\"initialize":"","finalize":"","x":260,\"y":160,\"wires":[["432218e0.808c58"]]}, {"id":"432218e0.808c58","type":"http request","z":"21197ab8.4d52c6","name":"","method":"POST","ret":"obj","paytoqs":"ignore","url":"https://iam.cloud.ibm.com/identity/token","tls":"","persist":false,"proxy":"","authType":"","x":412.00000381469727,\"y\":204.00000286102295,\"wires":[["cbd7d859.081478"],"4daad19.5978f3"]]}, {"id":"cbd7d859.081478","type":"debug","z":"21197ab8.4d52c6","name":"","active":true,\"tosidebar\":true,\"console\":false,\"tostatus\":false,\"complete\":\"payload\",\"targetType\":\"msg\",\"statusVal":"","statusType\":\"auto\",\"x\":600,\"y\":260,\"wires":[]}, {"id":"4daad19.5978f3","type":"function"}]
```

```

n","z":"21197ab8.4d52c6","name":"Pre Prediction","func":"var ag = global.get(\"ag\")\nvar sx =
global.get(\"sx\")\nvar bmi = global.get(\"bmi\")\nvar ch = global.get(\"ch\")\nvar sm =
global.get(\"sm\")\nvar rg = global.get(\"rg\")\nvar
token=msg.payload.access_token\nmsg.headers={'Content-Type':
'application/json','Authorization':'Bearer
'+token,'Accept':'application/json'}\nmsg.payload={\"input_data\":{\"fields\": [[\"Age\",
\"Gender\", \"BMI\", \"Children\", \"Smoker\", \"Region\"]],\"values\":
[[ag,sx,bmi,ch,sm,rg]]}}\nreturn
msg;\",\"outputs\":1,\"noerr\":0,\"initialize\":\"\",\"finalize\":\"\",\"x\":604.0000076293945,\"y\":156.000002
1457672,\"wires\":[[\"66912953.d1e5f8\"]],{\"id\":\"66912953.d1e5f8\",\"type\":\"http
request\",\"z\":\"21197ab8.4d52c6\",\"name\":\"\",\"method\":\"POST\",\"ret\":\"obj\",\"paytoqs\":\"ignore\",\"ur
l\":\"https://us-south.ml.cloud.ibm.com/ml/v4/deployments/ca4620cc-04cb-49ba-a8a9-
e8f96c92165e/predictions?version=2020-09-
01\",\"tls\":\"\",\"persist\":false,\"proxy\":\"\",\"authType\":\"\",\"x\":810,\"y\":200,\"wires\":[[\"8e571fd8.aafe3\",
\"b42674f6.1d7948\"]],{\"id\":\"8e571fd8.aafe3\",\"type\":\"debug\",\"z\":\"21197ab8.4d52c6\",\"name\":\"
\",\"active\":true,\"tosidebar\":true,\"console\":false,\"tostatus\":false,\"complete\":\"false\",\"statusVal\":\"\",
\"statusType\":\"auto\",\"x\":982.9999914169312,\"y\":264.00000381469727,\"wires\":[]},{\"id\":\"9bbac
2e7.5d459\",\"type\":\"inject\",\"z\":\"21197ab8.4d52c6\",\"name\":\"\",\"props\":{\"p\":\"payload\"},{\"p\":\"to
pic\",\"vt\":\"str\"}],\"repeat\":\"\",\"crontab\":\"\",\"once\":false,\"onceDelay\":0.1,\"topic\":\"\",\"payload\":\"\",
\"payloadType\":\"date\",\"x\":130,\"y\":60,\"wires\":[[\"a49c4564.85e1b8\"]],{\"id\":\"55741e81.de1eb\",\"ty
pe\":\"debug\",\"z\":\"21197ab8.4d52c6\",\"name\":\"\",\"active\":true,\"tosidebar\":true,\"console\":false,\"tos
tatus\":false,\"complete\":\"false\",\"statusVal\":\"\",\"statusType\":\"auto\",\"x\":813.9999885559082,\"y\":4
7.0000057220459,\"wires\":[]},{\"id\":\"b42674f6.1d7948\",\"type\":\"function\",\"z\":\"21197ab8.4d52c6
\",\"name\":\"Parsing\",\"func\":\"msg.payload=msg.payload.predictions[0].values[0][0]\nreturn
msg;\",\"outputs\":1,\"noerr\":0,\"initialize\":\"\",\"finalize\":\"\",\"x\":659.6000595092773,\"y\":88.0000019
0734863,\"wires\":[[\"55741e81.de1eb\",\"8d57eed8.7e906\"]],{\"id\":\"8d57eed8.7e906\",\"type\":\"ui_t
ext\",\"z\":\"21197ab8.4d52c6\",\"group\":\"8e58286.aca2fd8\",\"order\":2,\"width\":0,\"height\":0,\"name\":
\"\",\"label\":\"Prediction\",\"format\":{\"msg.payload}}\",\"layout\":\"row-
spread\",\"x\":777.6000366210938,\"y\":321.20001220703125,\"wires\":[]},{\"id\":\"8e58286.aca2fd8\",
\"type\":\"ui_group\",\"z\":\"\",\"name\":\"New\",\"tab\":\"73a4dea6.3844c\",\"order\":1,\"disp\":true,\"width\":
\"6\",\"collapse\":false},{\"id\":\"73a4dea6.3844c\",\"type\":\"ui_tab\",\"z\":\"\",\"name\":\"Home\",\"icon\":\"dash
board\",\"disabled\":false,\"hidden\":false}}]

```