Institiúid Teicneolaíochta Cheatharlach
INSTITUTE of TECHNOLOGY CARLOW
At the Heart of South Leinster

Machine learning model to categorize physical activity based on job description

Nithin Kumar Nukala **Supervisor: Dr. Oisin Cawley**

Introduction

For the first time in 2018 insurance premiums market over the world crossed $5 trillion mark. In order to attract the customers insurance companies has come up with many policies in both life insurance and non life insurance sectors. In life insurance sector the policy depends on various parameters like age, salary, occupation and marital status. The claim process and duration also varies based on the type of the policy.

An important indicator of potential claims duration is the type of job in terms of physical activity such as sedentary, light, medium or heavy. Based on this category duration of the claim varies. It would be really helpful if there is a model that can automate the manual process of assigning.

Dataset

Data for this model is provided by Unum insurance company. It is one of the fortune 500 company. Additional data is taken from U.S. Bureau of Labor Statistics.

Data contains 1 million records of job type and relevant information.

Literature Review

An important feature of technology is it generates large amounts of digital data. End goal for a company in hiring data scientist is for decision making meaning they need to analyse the data and find insights that helps companies to make proper decisions.

Clustering is an unsupervised model used for partitioning data. Already using in the insurance companies for segmentation of policies and potential buyers. Risk classification, fraud detection is also implemented. K-means clustering is used to find high profit customers.

Research Objective

- Can we automate the assignment of the physical activity category based on a combination of job title description and relevant external data with little or no manual intervention?
- Create the accurate model that can assign the category off the assignment.

Year	Life	Nonlife (2)	Total
2016	$2,576,886	$2,117,918	$4,694,804
2017	2,724,017	2,233,490	4,957,507
2018	2,820,175	2,373,050	5,193,225

Figure 1: World insurance premiums in dollars

Methodology

- All the jobs present in the Unum data should be processed in order to modify them into appropriate jobs using lemmatisation and stemming techniques.
- Next step would be mapping the Unum jobs with the relevant jobs from standard U.S. SOC(Standard Occupational Classification) job titles. This can be done using fuzzy logic mapping technique.
- Finally apply clustering algorithms and create 4 categories such as sedentary, light, medium and heavy based on physical activity of person

Technologies

scikit learn

python

teradata.

Future work

- We can further micro segmentation on the insurance data in order to find potential buyers.
- We can also use this clustering technique to create new policies based on buyer data

References:

1. bls(2018). U.S. Bureau of labour statistics. [online] Available at: https://www.bls.gov/soc/2018/home.htm [Accessed 7 Apr. 2020]
2. Iii. Insurance Information Institute. [online] Available at: https://www.iii.org/publications/insurance-handbook/economic-and-financial-data/world-insurance-marketplace [Accessed 7Apr. 2020]
3. Ma Hong, Kang Jing and Liu Li-xiong (2010) 'Research on clustering algoriths of data streams', 2010 2nd IEEE International Conference on Information Management and Engineering, Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on, pp. 1–4. doi: 10.1109/ICIME.2010.5477935