Insurance product purchase prediction using Artificial Neural Networks

Objective of research

The main objective of this research is to determine how well artificial neural networks perform in predicting whether customers purchase a certain insurance product. This analysis will give the insurance company a way to understand the purchasing behavior of its customers and predict it.

Data description

The dataset consists of seven variables/columns. The variables are as follows:

- Age - age of the customer
- Sex – gender of the insurance customer, male/female
- Years since joining – The number of years the individual has been a customer of this insurance company
- Marital status – The marital status of the customer
- Occupation category – The category of the customer's occupation
- Class – whether or not the customer purchased the product (0-No, 1-Yes)

```
data.head()
```

	years_since_joining	sex	marital_status	age	occupation_category	class
0	1.00	F	M	33	T4MS	0
1	0.99	F	M	39	T4MS	0
2	6.99	M	U	29	90QI	0
3	0.98	M	M	30	56SI	0
4	0.98	M	M	30	T4MS	0

```
data.describe()
```

	years_since_joining	age	class
count	29131.000000	29131.000000	29131.000000
mean	2.275411	40.482991	0.234561
std	2.145143	9.325760	0.423731
min	0.000000	9.000000	0.000000
25%	0.990000	33.000000	0.000000
50%	1.980000	40.000000	0.000000
75%	2.980000	47.000000	0.000000
max	120.000000	88.000000	1.000000

Data cleaning and feature engineering

All the variables were checked to ensure that there were no missing values. There were indeed no missing values in the data.

The categorical variables sex, marital status and occupation category were converted to numerical using Pandas functions. This is so that we can fit them well in our regression models.

Training the models models

Three Artificial Neural network based models were trained on the data using 50 epochs and their performances were measured and compared based on the roc-auc score as shown in the table below. The models all had one hidden layer and the hidden layers had 7, 10 and 20 nodes respectively.

Number of nodes in hidden layer	Roc-auc score
7	0.591
10	0.561
20	0.561

Choice of final model

The best model is the one that achieved the highest roc-auc score which is the model with 7 nodes in the hidden layer and this is our final model.

Summary of key findings

The models took a long time to train and they were not able to attain a very great roc-auc score. In conclusion artificial neural networks may not be very suitable for this problem.

Suggestion for next steps

We should attempt to solve this problem using other classification models like SVM or Random Forest as these may achieve better results compared to Artificial Neural Networks.