An analysis of the factors that determine whether a customer purchases a certain insurance product

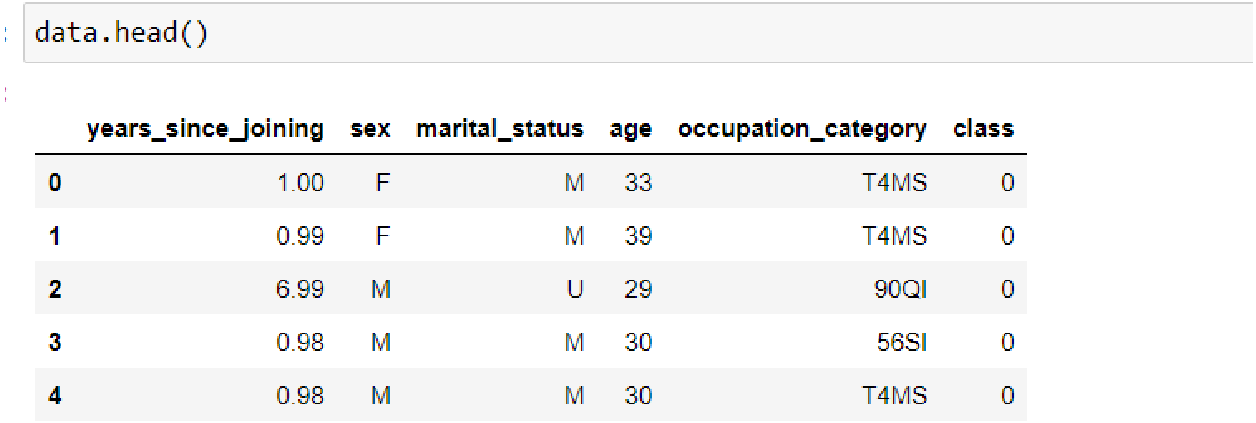
Objective of research

The main objective of this research is to determine the main factors that determine whether a customer purchases a certain insurance product. This study focuses on interpretation of the factors that determine whether a customer purchases the insurance product. This analysis will give the insurance company a way to understand the purchasing behavior of its customers.

Data description

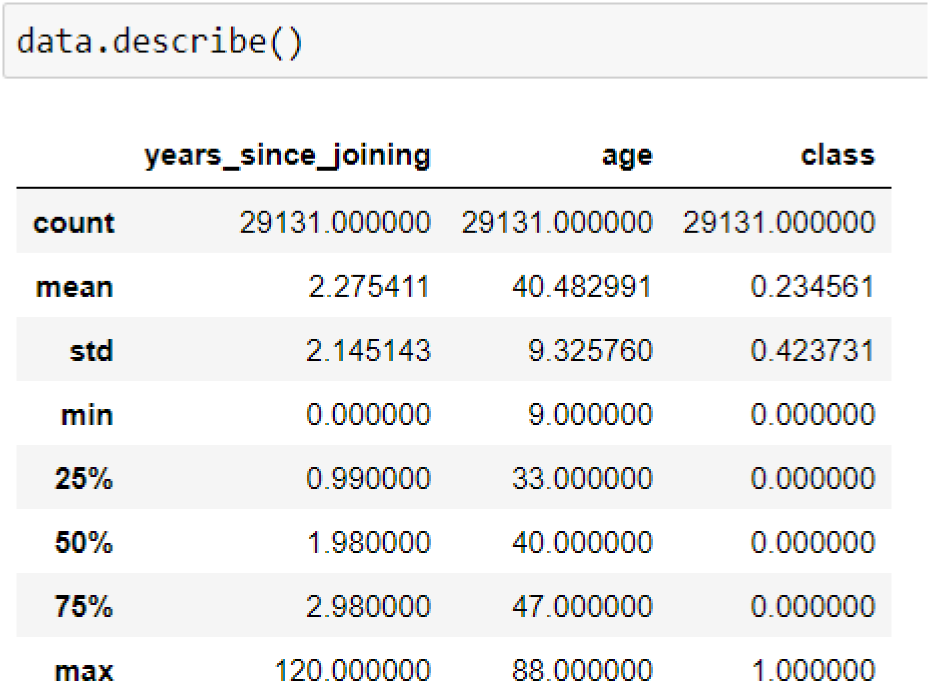
The dataset consists of seven variables/columns. The variables are as follows:

- 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

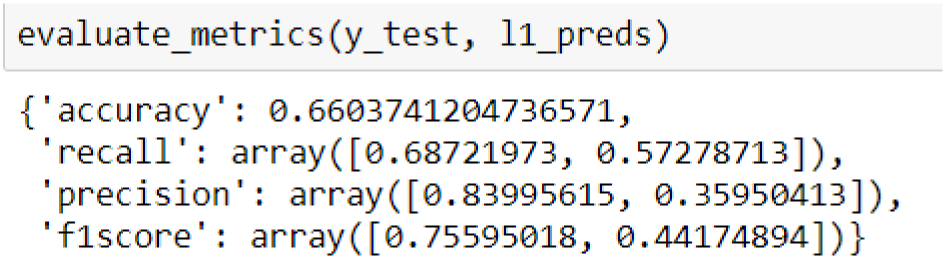
All the variables were checked to ensure that there were no missing values. There were indeed no missing values in the data.

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 classification models models

Logistic regression

Firstly, a simple logistic regression model with L1 penalization was fitted to the data. The classes were imbalanced so class reweighting was used. The performance of the model is shown in figure 1 below.



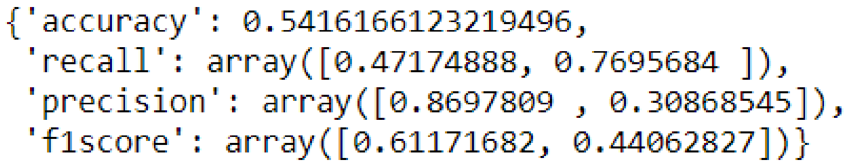
```
evaluate_metrics(y_test, l1_preds)
```

```
{'accuracy': 0.6603741204736571,
 'recall': array([0.68721973, 0.57278713]),
 'precision': array([0.83995615, 0.35950413]),
 'f1score': array([0.75595018, 0.44174894])}
```

Figure 1: Performance of logistic regression model

Support Vector Classifier

A support vector classifier was fitted to the data. The classes were imbalanced so class reweighting was used. The performance of the model is shown in figure 2 below.

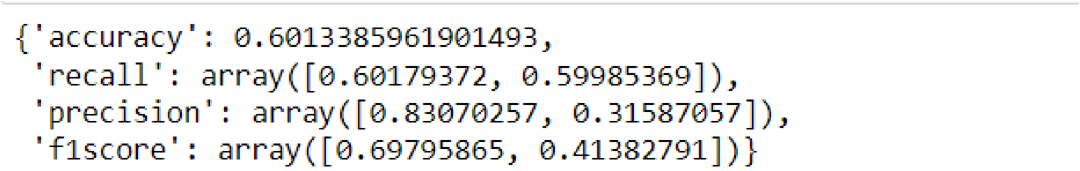


```
{'accuracy': 0.5416166123219496,
 'recall': array([0.47174888, 0.7695684 ]),
 'precision': array([0.8697809 , 0.30868545]),
 'f1score': array([0.61171682, 0.44062827])}
```

Figure 2: SVM classifier performance

Random Forest Classifier

A random forest classifier was fitted to the data. The classes were imbalanced so class reweighting was used. The performance of the model is shown in figure 2 below.



```
{'accuracy': 0.6013385961901493,
 'recall': array([0.60179372, 0.59985369]),
 'precision': array([0.83070257, 0.31587057]),
 'f1score': array([0.69795865, 0.41382791])}
```

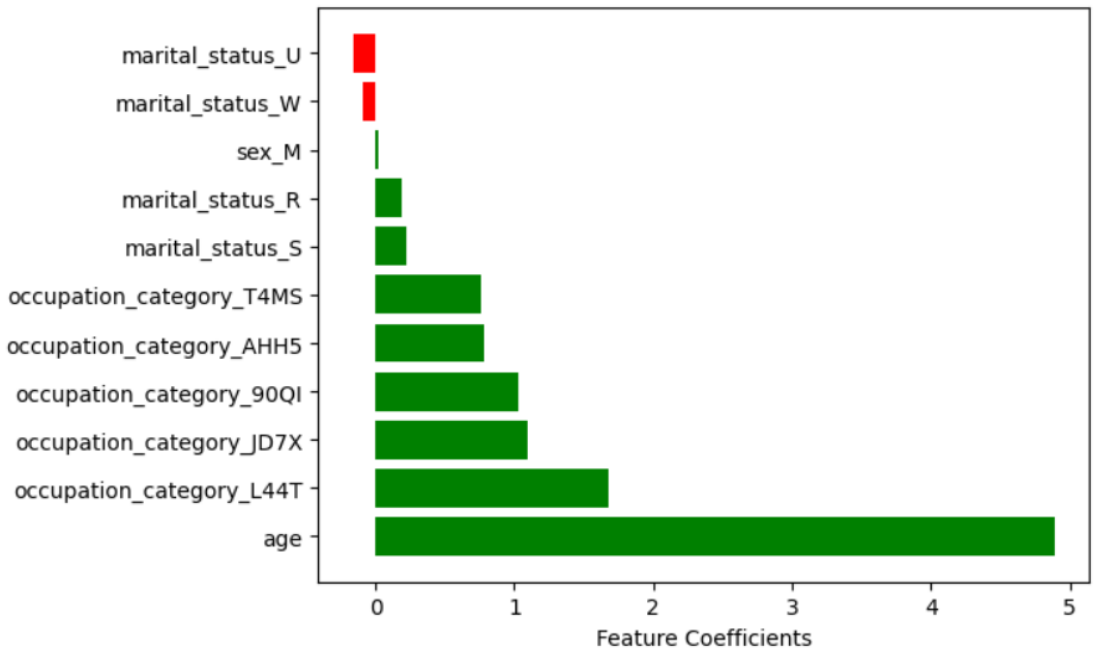
Figure 3: Random Forest Classifier performance

Choice of final model

The logistic regression model had the best performance in terms of accuracy and recall within both classes. Furthermore, it is the model which is most easy to explain between the three models. The final model chosen is therefore the logistic regression model.

Summary of key findings

The key findings from this research are how much each of the factors determine whether aa customer purchases an insurance product. Figure 4 below shows how much each factor influences the odds of a customer purchasing the insurance product.



From the graph we see that age has most influence on the odds of a customer purchasing the insurance product followed by occupation category etc. Marital status U and W are the only features with negative coefficients so if a customer has marital status U or W it reduces their odds of purchasing the insurance product.

Suggestion for next steps

All the three models were not able to attain a good balance between accuracy, precision and recall for the two classes being modelled. This might be because the classes were heavily imbalanced and the research should be repeated and down sampling or up sampling techniques included.

Another reason might be that the available explanatory variables were not enough to explain the process. More explanatory variables might be needed to try and explain this purchasing process by customers.