School of Engineering and Applied Science
CSE523- Machine Learning

Project Report - 2

Vehicle Insurance Predictor

Submitted to: **Prof. Mehul Raval**
Date of submission**: 16th Feb 2021**

Members:

Name	Roll Numbers	Branch
Samarth Shah	AU1841145	B. Tech, ICT
Priyank Sangani	AU1841136	B. Tech, ICT
Yash Patel	AU1841141	B. Tech, ICT
Shaili Gandhi	AU1841012	B. Tech, ICT

Task performed this week:

- Background research for the selected topic
- Preparing the categorical columns
- Converting categorical columns into dummies and getting converted into numerical values
- Understanding data by getting a description of data and correlation matrix between the data columns.
- Also eliminating the null values of the columns and eliminating columns that are not needed (eg. id column).

Outcomes of the tasks performed:

- Dummy columns: 1) Vehicle Age, 2) Is_damaged, 3) Gender

	1-2 Year	< 1 Year	> 2 Years
0	0	0	1
1	1	0	0
2	0	0	1
3	0	1	0
4	0	1	0

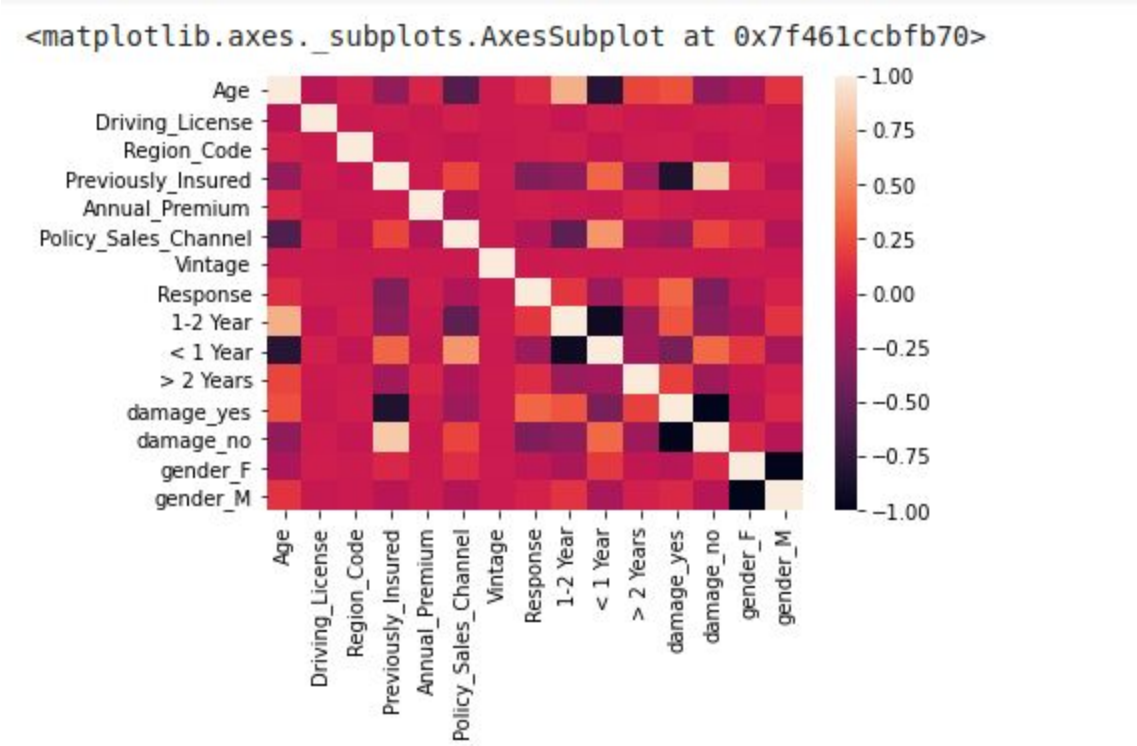
	No	Yes
0	0	1
1	1	0
2	0	1
3	1	0
4	1	0

	Female	Male
0	0	1
1	0	1
2	0	1
3	0	1
4	1	0

- Correlation matrix:

	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response	1-2 Year	< 1 Year	> 2 Years	damage_yes	damage_no	gender_F	gender_M
Age	1.000000	-0.079782	0.042574	-0.254682	0.067507	-0.577826	-0.001264	0.111147	0.692910	-0.787775	0.220694	0.267534	-0.267534	-0.145545	0.145545
Driving_License	-0.079782	1.000000	-0.001081	0.014969	-0.011906	0.043731	-0.000848	0.010155	-0.037403	0.040215	-0.006211	-0.016622	0.016622	0.018374	-0.018374
Region_Code	0.042574	-0.001081	1.000000	-0.024659	-0.010588	-0.042420	-0.002750	0.010570	0.038055	-0.044250	0.014555	0.028235	-0.028235	-0.000604	0.000604
Previously_Insured	-0.254682	0.014969	-0.024659	1.000000	0.004269	0.219381	0.002537	-0.341170	-0.279077	0.358773	-0.191352	-0.824143	0.824143	0.081932	-0.081932
Annual_Premium	0.067507	-0.011906	-0.010588	0.004269	1.000000	-0.113247	-0.000608	0.022575	-0.002495	-0.022555	0.061918	0.009349	-0.009349	-0.003673	0.003673
Policy_Sales_Channel	-0.577826	0.043731	-0.042420	0.219381	-0.113247	1.000000	0.000002	-0.139042	-0.508265	0.571516	-0.146238	-0.224377	0.224377	0.111159	-0.111159
Vintage	-0.001264	-0.000848	-0.002750	0.002537	-0.000608	0.000002	1.000000	-0.001050	-0.002632	0.002410	0.000600	-0.002064	0.002064	0.002517	-0.002517
Response	0.111147	0.010155	0.010570	-0.341170	0.022575	-0.139042	-0.001050	1.000000	0.164317	-0.209878	0.109300	0.354400	-0.354400	-0.052440	0.052440
1-2 Year	0.692910	-0.037403	0.038055	-0.279077	-0.002495	-0.508265	-0.002632	0.164317	1.000000	-0.918704	-0.220402	0.284717	-0.284717	-0.147633	0.147633
< 1 Year	-0.787775	0.040215	-0.044250	0.358773	-0.022555	0.571516	0.002410	-0.209878	-0.918704	1.000000	-0.182750	-0.370778	0.370778	0.166280	-0.166280
> 2 Years	0.220694	-0.006211	0.014555	-0.191352	0.061918	-0.146238	0.000600	0.109300	-0.220402	-0.182750	1.000000	0.206961	-0.206961	-0.043155	0.043155
damage_yes	0.267534	-0.016622	0.028235	-0.824143	0.009349	-0.224377	-0.002064	0.354400	0.284717	-0.370778	0.206961	1.000000	-1.000000	-0.091606	0.091606
damage_no	-0.267534	0.016622	-0.028235	0.824143	-0.009349	0.224377	0.002064	-0.354400	-0.284717	0.370778	-0.206961	-1.000000	1.000000	0.091606	-0.091606
gender_F	-0.145545	0.018374	-0.000604	0.081932	-0.003673	0.111159	0.002517	-0.052440	-0.147633	0.166280	-0.043155	-0.091606	0.091606	1.000000	-1.000000
gender_M	0.145545	-0.018374	0.000604	-0.081932	0.003673	-0.111159	-0.002517	0.052440	0.147633	-0.166280	0.043155	0.091606	-0.091606	-1.000000	1.000000

- Correlation matrix heat map:



- Data description:

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
count	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000
mean	190555.000000	38.822584	0.997869	26.388807	0.458210	30564.389581	112.034295	154.347397	0.122563
std	110016.836208	15.511611	0.046110	13.229888	0.498251	17213.155057	54.203995	83.671304	0.327936
min	1.000000	20.000000	0.000000	0.000000	0.000000	2630.000000	1.000000	10.000000	0.000000
25%	95278.000000	25.000000	1.000000	15.000000	0.000000	24405.000000	29.000000	82.000000	0.000000
50%	190555.000000	36.000000	1.000000	28.000000	0.000000	31669.000000	133.000000	154.000000	0.000000
75%	285832.000000	49.000000	1.000000	35.000000	1.000000	39400.000000	152.000000	227.000000	0.000000
max	381109.000000	85.000000	1.000000	52.000000	1.000000	540165.000000	163.000000	299.000000	1.000000

Tasks to be performed in the upcoming week:

- We want to explore the topic a little more and get to define the reason why this topic is important.
- We would like to explore the dataset first and as we have done with data cleaning and processing so we would like to get the insights behind each and every feature and get some inferences like how they are correlated internally and how they behave with the label variable.