Data Analyst certification

Author: Dmitry Moskalev

Table of contents

Input data

Column name	Details
Age	Numeric, the customer's age
Employment Type	Character, the sector of employment
GraduateOrNot	Character, whether the customer is a college graduate
AnnualIncome	Numeric, the customer's yearly income
FamilyMembers	Numeric, the number of family members living with the customer
ChronicDiseases	Numeric, whether the customer has any chronic conditions
FrequentFlyer	Character, whether a customer books frequent tickets
EverTravelledAbroad	Character, has the customer ever travelled abroad
TravelInsurance	Numeric, whether the customer bought travel insurance

Travel Assured company wants to get some useful insights of the marketing strategy.

Input data contains in "*travel_insurance.csv*" file with **9 columns (features)** and **1987 rows.**
Items colors:

- The "*green*" items do not require changes
- The "*yellow*" items require changes with concatenating names for possibility of next data processing
- The "*orange*" items require changes due to type incompatibility with "Numeric" one

Example of raw (input) data

Age	Employment Type	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
31	Government Sector	Yes	400000	6	1	No	No	0

Example of data for processing

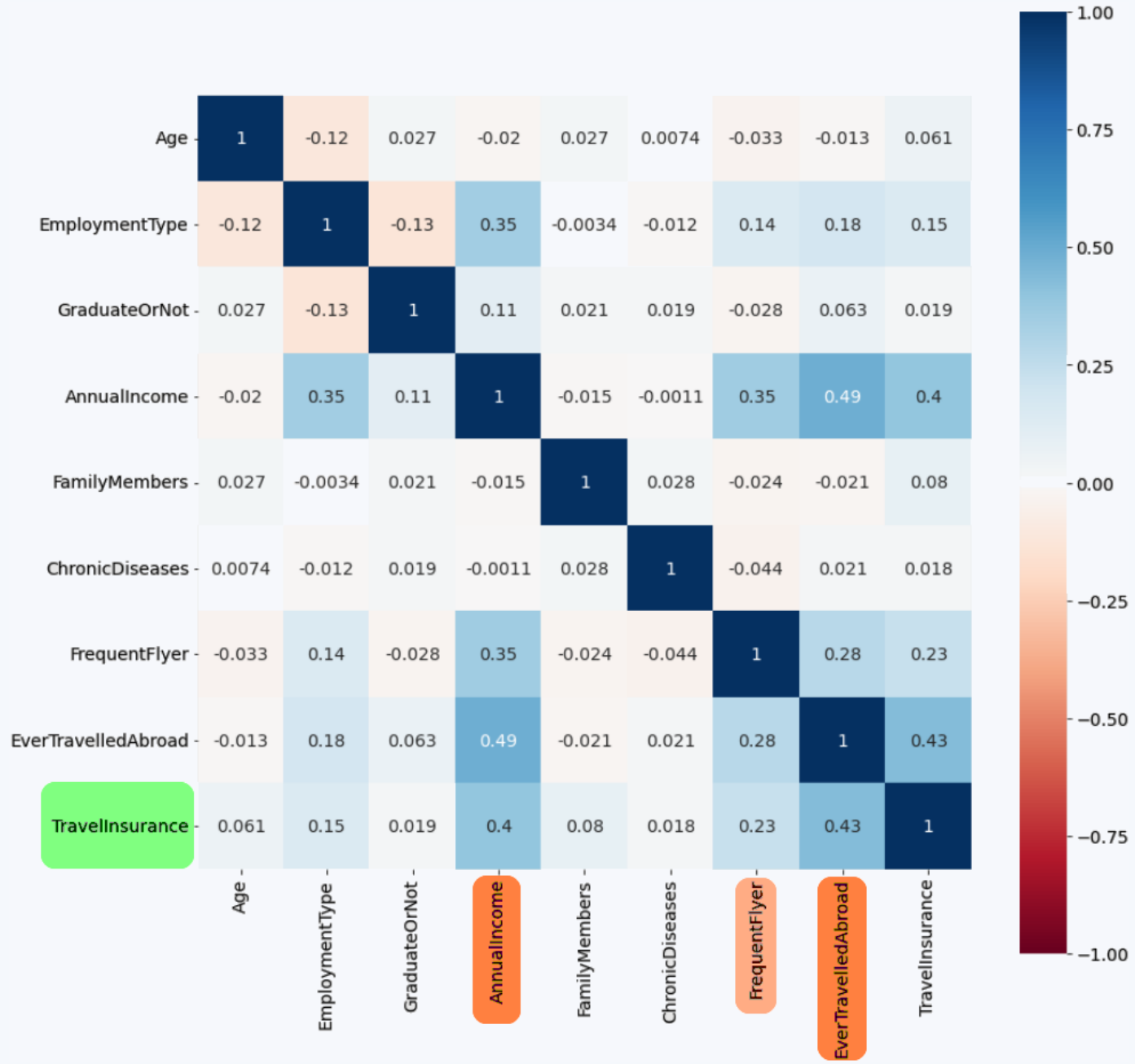
Age	EmploymentType	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
31	0	1	400000	6	1	0	0	0

Data analysis

Let's look to the target feature "***TravelInsurance***" in current report using the Pearson's correlation matrix.

Coefficients with value greater that zero indicates direct dependence and values less that zero indicates reverse dependence, zero value means missing of any dependence between features.
When dependence is direct the greater X values is, the greater Y values will be.
Reverse dependence means for a larger X values match a smaller Y values.

The greatest absolute value for "***TravelInsurance***" correlated with next features:
"***FrequentFlyer***" with 0.23 value;
"***AnnualIncome***" with 0.4 value;
"***EverTravelledAbroad***" with 0.43 value.
It means, that these 3 columns are also target features, because of making the most feature importance for clients who choosing to buy travel insurance or not.

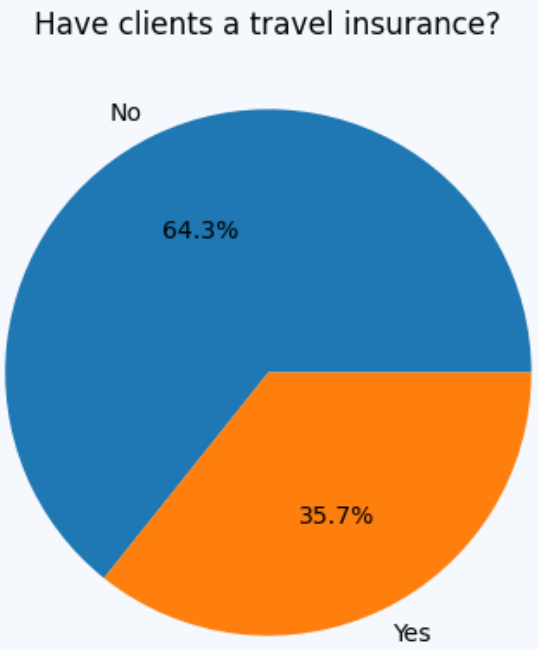


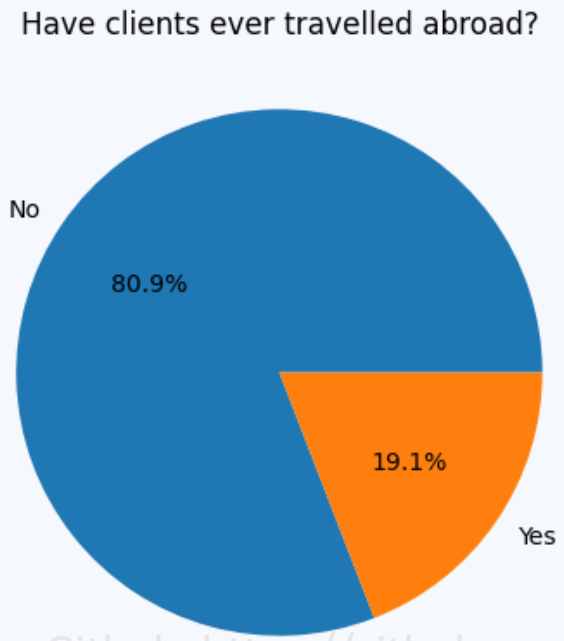
Data analysis

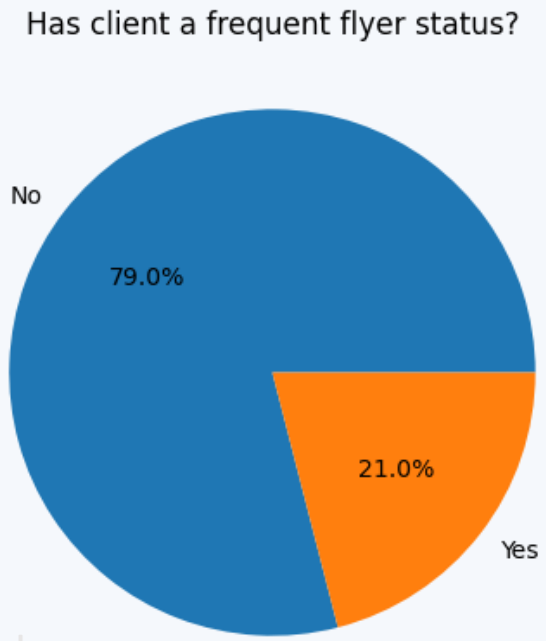
Let’s look to the minimum, maximum, average and difference of numeric values for better understanding.

Values	Age	EmploymentType	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
Minimum	25	0	0	300000	2	1	0	0	0
Maximum	35	1	1	1800000	9	0	0	0	0
Average	~29			~932762	~5				
Difference	10			1500000	7				

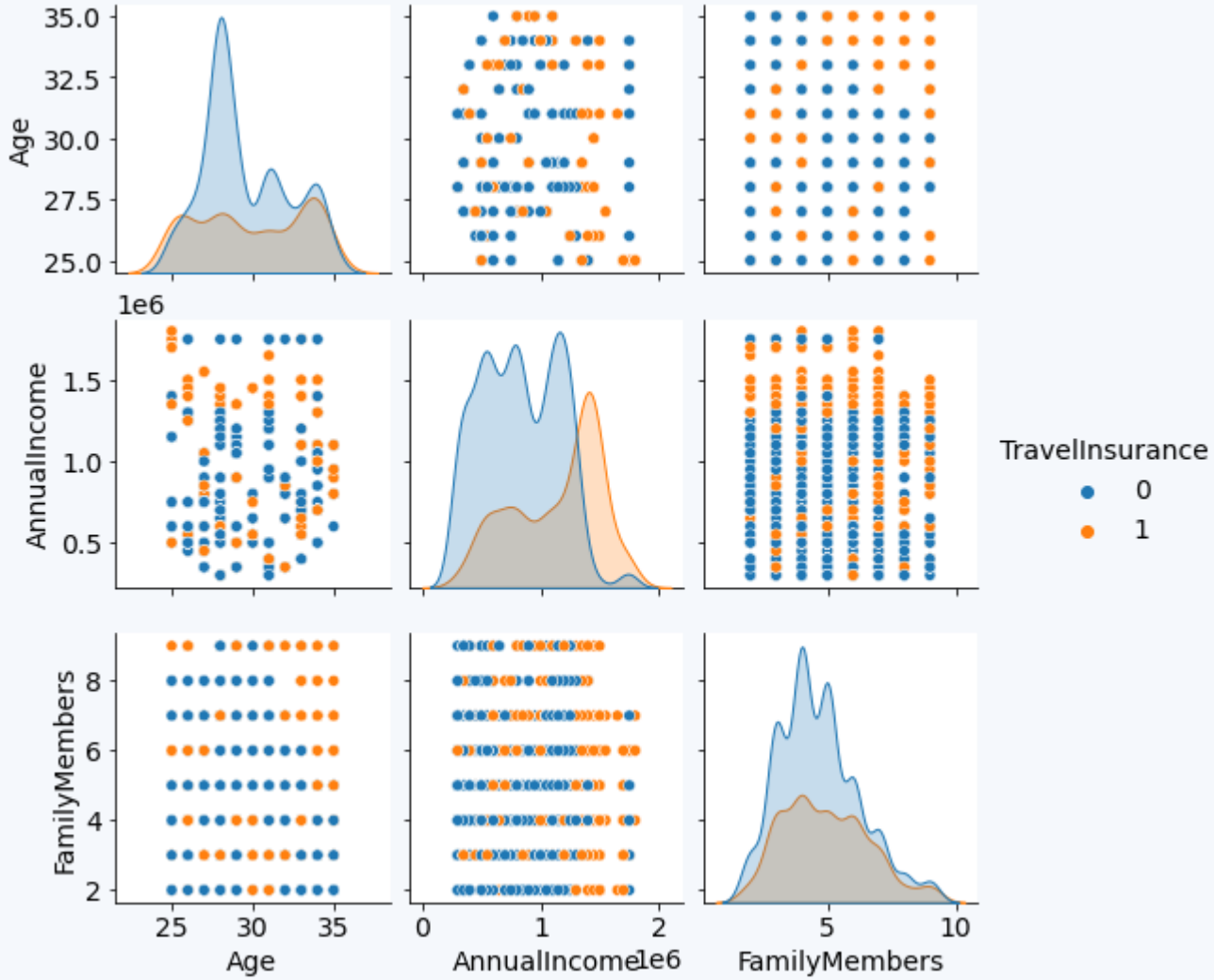
Class balance on the example of comparison for a few binary target features.

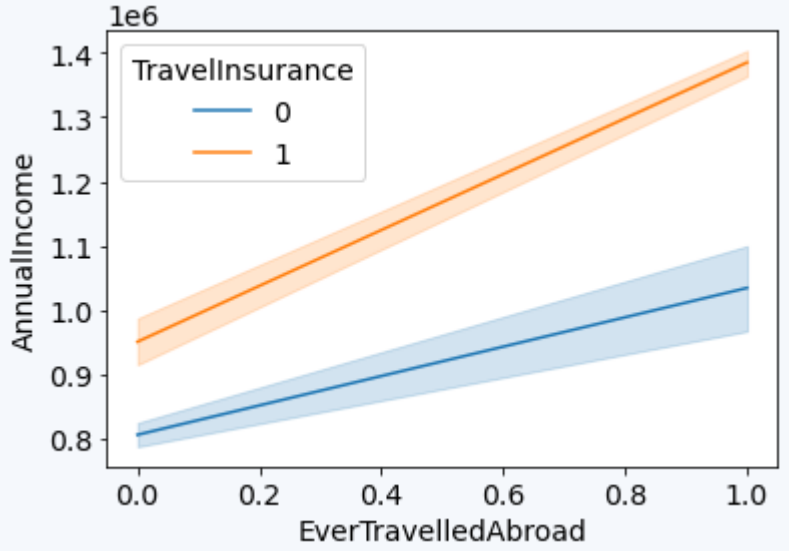


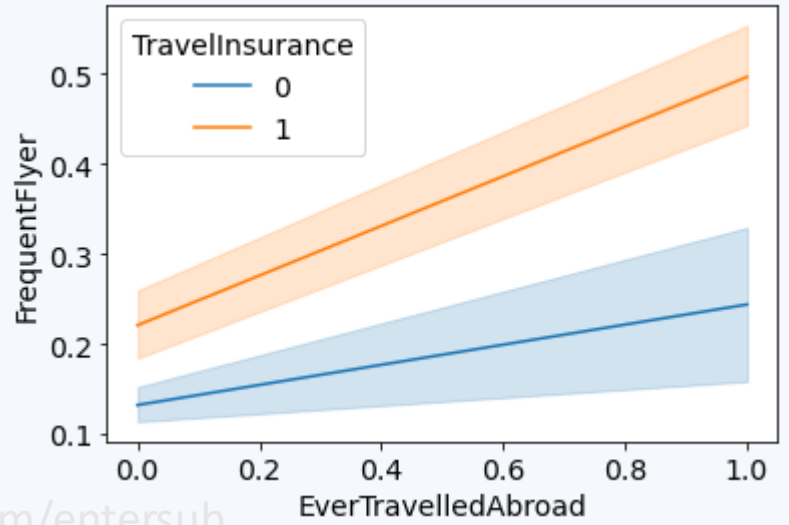




Data analysis







- The "*TravelInsurance*" feature value of 1 means that customer bought a travel insurance, 0 value means not
- Clients, who have traveled abroad at least once, customers with more annual income and clients with status of frequent flyer are taking travel insurance more than other

Results

The main **insights** of differences in the travel habits between a customers with a travel insurance instead of a customers without it are:

- *More than average annual income*
- *Have travelled abroad at least once*
- *Frequently flyers passengers*

In this case, the **target audience** is the customers younger than 28 years old with more than average annual income with possibilities of taking non-mandatory costs as travel insurance.

The **marketing strategy** is to advertise for these customers a travel insurance with more amount of benefits than competitors or with a lower price. The measure of success can be calculated as a profit by multiplying amount of sold travel insurance with pricing of each one.

Thanks for your attention!