Data Science Interview Task: Predicting Insurance Risk

A Jupyter Notebook that demonstrates how to define a problem statement, synthesise data, and apply a machine learning service

We have a theoretical problem for you to complete that should allow you to showcase your Data Science skills and commercial knowledge.

# Problem Statement

We have been approached by the business to use machine learning to determine risk. There are people that are dubious about the use of modern machine learning techniques over traditional risk quantification methods. We need to showcase how and why modern machine learning should be adopted, and the modelling approach and outputs should be easily explainable to stakeholders.

The goal of this task is to predict the risk of offering an insurance policy to a potential customer, based on their personal and financial information. The risk is defined as the probability of the customer filing a claim within a year of purchasing the policy. A higher risk means a higher chance of the customer making a claim, which would result in a loss for the insurance company. Therefore, the insurance company wants to understand the risk profile of each customer to determine how to price products.

The business would like to understand how we could use this model to ensure that only 5% of the people they sell a policy to would claim on their policy.

The first step is to produce a Jupyter notebook that is completely self-contained and explains the problem, the approach and a solution. The person reviewing the notebook has no knowledge of the context of this piece of work and will only have the notebook to go on. You can assume they have a good understanding of python and Jupyter notebooks.

# Data Synthesis

To demonstrate the machine learning service, we need you to synthesise some data that resembles the real-world data that an insurance company would have. You will need to generate a dataset of 10000 customers, each with 40 features that describe their personal and financial characteristics. Some of the features could be numerical, such as age, income, credit score, and number of dependents. Some of the features could be categorical, such as gender, marital status, education level, and occupation. Some of the features will be ordinal, such as health status, driving record, and home ownership. Finally, we would expect to have some missing values and duplicates, if these were real world data. You can either synthesise this in, and it handle later, or discuss how you would have treated the data if it were there.

The target variable is the binary claim status of 0 or 1.

# Data Exploration and Preprocessing

We want to understand your process for tackling the task, i.e. how the data is explored and manipulated is how we understand your process. What assumptions are you making and why are you using the approach you have gone with?

# Modelling

Now that we have a clean and ready dataset, we can make a model and see how it performed. How will you present the performance of the model and explain its good and bad points. In the insurance world, explainability is paramount, so the business has to be comfortable that they understand what the model is doing. NOTE: It does not matter how good / poor the actual model performance is as this is only a synthesised dataset and a proof of concept.

# Productionising

Assuming the business likes your model, what are the next steps to go from a Jupyter Notebook to a productionised service?

What are the considerations for ensuring the business can leverage it?

What are the steps you would take to provide this to the business?

Which traditional teams in a business would you need to speak to?

What is in scope and out of scope for your responsibility?

# Deliverables

Please send us your self-contained Jupyter Notebook for us to review. You can include answers to questions in the notebook or a separate document.