GemSeek

Data Scientist Evaluation

Preliminary Information

This document contains list of the tasks we would like you to complete in order to evaluate fully your technical abilities. Please read thoroughly all tasks and in case questions arise, do not hesitate to ask us.

You have 5 working days to complete all the tasks. However, if by some reason you cannot get all results in full, don't worry. The purpose of this exercise is not only to get exact numbers, but also to understand how you would approach similar assignments, what would be your way of thinking and how you would advise your Client at the end.

You are to work on a few separate tasks. Each requires some data handling, modelling or visualizations. We would like to get following outputs from you:

- All the pseudo-code / actual code from your software of preference. Please don't forget to include notes and comments within.
- Numerical outputs of the final models and statistical calculations (if applicable for the specific task)
- Data files with the required outputs in csv and/or .xslx formats
- Any visualizations of the data you consider useful to support your work (interactive charts would be considered as an advantage)
- Any other materials that you consider useful would be welcomed as well

At the end of the period you would be required to preset your work in front of the team or via Skype.

Many thanks and good luck!

The GemSeek's Data Science Team

Task1: Data Prep

You were supplied with a data set “test_sales.csv”. You have been told that the file contains the sales data for a specific client “X”

Task 1.1:
Create a dataset (another csv file) that contains for every order the count and the sum of all orders for the same customer in the past 3 months, as well as columns that contain the count and sum for the future 3 months of orders, again for any given customer. All of the derived fields should not include the observation point (order)
The produced dataset should have the following column structure:
1) order_numer
2) numer_of_orders_past_3_months
3) gross_sum_orders_past_3_months
4) numer_of_orders_future_3_months
5) gross_sum_orders_future_3_months

The produced data frame should be unique by order number.

Task 1.2. :
Aggregate the difference in overall historical activity and the future activity using the produced data frame in Task 1 and add a short comment.
Bonus: Derive the same statistics by “channel” type.

Task 2: KDA and prediction

Given the house price data (‘*housing_price_data.rar*’), predict the sale price (‘*SalePrice*’) based on the other house features in the dataset using linear regression.

Do the same with a modeling approach of your choice.

Compare and present results. Provide arguments for pros and cons of different methods, argument your choice from other similar

Task 3: Dimension reduction and aggregation

On the dataset 'dimension_data.csv' perform the following tasks:

1. Perform a factor analysis on all variables except *'respid'* and *'On.a.scale.of.zero.to.ten..how.likely.are.you.to.recommend.our.services.to.a.friend.or.colleague'*. Give the factors meaningful names and interpretations based on their loadings.

2. Assess the variable importance of the variables in the previous task relative to the variable: '*On.a.scale.of.zero.to.ten..how.likely.are.you.to.recommend.our.services.to.a.friend.or.colleague*'.