Data Science Challenge

Dear candidate,

You have made a good impression in the first steps in our recruitment process and we would like to progress with you to the next part of our assessment.
The purpose of the data science challenge is to let you demonstrate the way you think and work.
This exercise should take around 2 to 4 hours to be completed but you have 48 hours to return your solution.

Please take a look at the dataset in the file “Auto1-DS-TestData.csv” (see https://archive.ics.uci.edu/ml/datasets/Automobile for information on the features and other attributes) and answer the following questions:

Question 1 (10 Points)
List as many use cases for the dataset as possible.

As provided by https://archive.ics.uci.edu/ml/datasets/Automobile this dataset consists of 3 different entities:

a) the specification of an auto in terms of various characteristics
b) assigned insurance risk rating
c) normalized losses in use as compared to other cars.

Based on these entities I have thought of these use cases:

1- These features are really useful items which might be searched by the customers to target they desired car types. For example one is searching for a Cabrio BMW 320. So for getting the specific query via the search engine of AUTO1 these features might be really useful to find and list the cars which might be a potential candidate for the customer.

2- One could also found the relation between the features and extract the correlation between them. I mean for example for the users might be interesting that how much their desired car category burn fossil fuels. For example the mpg (miles per gallon) factor might be linked with weight, horse power, number of cylinder etc. Usually a customer might want to optimize her/his search based on many features and then decide what to buy. And in that case a kind of optimized recommendation system base on customers wishes might help them to decide faster and easier to land at their desired choice as the variety of cars are really large and confusing and the customer is never sure if she/he made the best choice at the end.

3- For the AUTO1 company it might be very useful to have an estimation of the risk they put in their business and if the return of the investment is some how guaranteed. It is where the insurance risk rating item would play a huge role. Based on the archived data one can build a statistical model to predict this quantity and assign this value on the candidate buys. This prediction if is done with an acceptable error range, will help to increase the benefits and reduce the loss.

4- Following the same strategy in 3, AUTO1 can create an automatic system to predict the real price of the used cars. This feature might be very interesting for the customers as they can easily take a fair and fast estimation of the price of their used car by applying this tools. It can also be offered as a free feature if the users sign into the AUTO1 web and share their data with the company and in return receive a fair and realistic estimate of their used cars price.

5- The other target variable which might be interesting for the company is the relative average loss payment per insured vehicle year, which represents the average loss per car per year. I think if the company has an estimate of the loss of the cars price with time after the buying, it can calculate the appropriate final price for selling.

Question 2 (10 Points)
Auto1 has a similar dataset (yet much larger...)
Pick one of the use cases you listed in question 1 and describe how building a statistical model based on the dataset could best be used to improve Auto1’s business.

I will go for the normalized loss ratio or option 5, which I think is very time dependent and will affect the company very largely in terms of the speed of selling the cars. The company is buying the cars and if the price is decreasing with time, it will put a real pressure in the selling process of the company or the partners.

Question 3 (20 Points)
Implement the model you described in question 2 in R or Python. The code has to retrieve the data, train and test a statistical model, and report relevant performance criteria.

When submitting the challenge, send us the link for a Git repository containing the code for analysis and the potential pre-processing steps you needed to apply to the dataset. You can use your own account at github.com or create a new one specifically for this challenge if you feel more comfortable.

Ideally, we should be able to replicate your analysis from your submitted source-code, so please explicit the versions of the tools and packages you are using (R, Python, etc).

Git : https://github.com/bijanfallah/A1_Bijan.git

Question 4 (60 Points)

A. Explain each and every of your design choices (e.g., preprocessing, model selection, hyper parameters, evaluation criteria). Compare and contrast your choices with alternative methodologies.

B. Describe how you would improve the model in Question 3 if you had more time.

A. In my code I have commented the steps I did for preprocessing. I have chosen 5 different models (Ridge, LASSO, Elastic Nets, Random Forest and XGBoost) to cover a good variety of linear and nonelinear models as well as ensemble methods. Before using the final models, I have trained them based on hyperparameters using gridsearchCV.

Finally, I tried to use stacking for my final model, which considers combination of the models as a final solution. For that I used 3 different methods: simple ensemble mean, linear averaging and Neural nets.

B. Well the model would be improved given large dataset. The observation count is very small for making the models more complex (over-fitting may happen here). And if a good large dataset is hand I would suggest to go for deeper neural nets. The tunning of the models in that stage might be also a challenge. Probably, there must also some dimensionality reduction methods be applied to reduce the volume of the data.

We are looking forward to receiving your solution!