ML Engineer case study

"We use Machine Learning and Data Science to make sport accessible to the many."

The general instructions are as follows:

● Please do the exercise using Python and your favorite IDE.. If you think you have better tools, we will be happy to hear.

● Export your work to some place where we can see your progress (your commits), for example Github. Please make sure you give viewing permission to the people on this email.

● Please make regular commits with your ongoing work. Do not upload the final result when you get it done. Instead, do your work in the public so we can follow your regular updates.

The questions are listed below. Please have a look, and do not hesitate to ask any question about the format of the data files, or the requirements of the questions.

Use case

Data

train.csv.gz & *test.csv.gz* files contain weekly store-department turnover data (*test.csv.gz* does not contain the turnover feature). For confidentiality reasons the turnover values are rescaled.
bu_feat.csv.gz file contains some useful store information.

1. Preliminary questions & EDA

a. Which department made the highest turnover in 2016 ?
b. What are the top 5 week numbers (1 to 53) for department 88 in 2015 in terms of turnover over all stores ?
c. What was the top performer store in 2014 ?
d. Based on sales can you guess what kind of sport represents department 73 ?
e. Based on sales can you guess what kind of sport represents department 117 ?
f. What other insights can you draw from the data? Provide plots and figures if needed. (Optional)

2. Modeling

In stores many decisions are made by managers at the department level.
In order to help store managers in making mid-term decisions driven by economic data, we want to forecast the turnover for the next 8 weeks at store-department level.

a. Build and evaluate performances of a simple estimator able to predict the turnover of test.csv.gz data.
b. The goal here is not to produce a state-of-the-art model of time series forecast.
c. Propose another model or strategy that may increase the quality of the predictions. (Optional)

3. ML pipeline

Based on the previous steps of data exploration and modeling and in order to deploy your model in production, develop a machine learning pipeline which:

a. Reads the raw data
b. Transforms the data in the right format for the model
c. Trains the model
d. Makes predictions and exposes the the results

The aim here is to realize a reproducible pipeline.

e. Describe some common issues involved in the deployment of machine learning models.
f. Propose a solution which monitors the the model performance in production (optional)