For this final project we were asked to work in a real-life project (available in public sources or not) and make use of the concepts learned and used during the semester. While it is a great way to deepen on this knowledge it also involves some research and learning on how to deal with a specific problem from the industry. Part of this project is a peer review on the project of another team. This peer review is for team Carpool who predict sales with data of the German drug company Rossmann.

Project Understanding

The challenge proposed by Rossmann Stores seems to be well understood by the team. A good explanation of the company's needs is given and also an explanation of the expected outcome.

However, a general part in the problem description about sales prediction is missing. It would be interesting to know if there are some success stories in the industry or in academic papers and what forecasting techniques they used.

In addition, the *head()* output for the train dataset doesn't add much useful information to the problem description; we think it's more adequate to include this in the data understanding section.

The link to the dataset/competition on Kaggle isn't working. By removing the final "s" in "stores" should work: https://www.kaggle.com/c/rossmann-store-sales

Data Understanding

With respect to the document organization, the Exploratory Data Analysis appears a bit late within the document. We think it would suit better immediately after the Problem Description section. Also the PCA analysis should be included as part of the Data Understanding.

The ratio of images is too high and some of them oversized. We recommend to use less images and explain the remaining ones a bit more.

Additionally, a description/explanation of each given feature is missing. What data do we have available?

The analysis of the sales by store type and day of week is great. However, it is not clear what day of the week you're considering as day 1. For example, if Monday is treated as the first day of the week you can reference the international standard ISO 8601. The same accounts for promo and day of week. One remark: we noticed that your promo variable is Boolean while the legend at the side of the plot shows a continuous Promo variable. By setting this variable as factor and assigning it to the *group* parameter in ggplot it can be solved.

The last two density graphs are explained with only one sentence each. And actually, does it add important information or it's superfluous? We noticed that according to the density plot the customers per day distributions seems to be very skewed to the right, are you going to mention something about this?

Since the PCA analysis isn't too meaningful for your project we suggest not to use for it that much space in the report. We have limited space so it's important to keep it for the most meaningful information.

Data Preparation

What do you mean by "the dataset was very clean"? Does it mean that there were no missing values, NAs, infinite values, etc.? Where there were no character variables? Outliers?

Are you planning to do some feature engineering or transform the data to get the most out of it for modeling? You mention that you joined the two training and store datasets. How many predictors do you have finally? We recommend to describe the joining process is more detail as it seemed to be a challenging task that you were able to manage.

Modeling

When you select lambda for your LASSO model, the support graphs are using log(lambda), but it's never mention/explained explicitly. We think it should be mentioned to make the report clearer.

How many folds are you using for cross-validation for LASSO? It isn't stated in the report either.

The explanation of the random forest model could be improved to reflect the concepts given in class. We recommend to use technical language instead of the non-technical wording in this paragraph ("a group of 'weak' models combine to form a more powerful model"). Also, you state that Random Forest “can handle thousands of input variables and identify the most significant variables. By doing so, it is a way of reducing dimensionality.” But is this even a relevant characteristic since you only have a few variables?

The report states that changing the seed for random numbers generation is part of the tuning of a model. To our understanding it is not. Setting a seed for random numbers generation is just a way to make sure that you will be able to get the same results every time you execute your program, even when make use of random numbers.

You didn't mention how you got the parameters for the Random Forest model. Did you do hyper-parameter tuning?

Evaluation

In the Kaggle competition it states that submissions are evaluated using the Root Mean Square Percentage Error (RMSPE) but in your report it looks like you're using Root Mean Square Error (RMSE). We suggest to use the same metric that Kaggle uses in order to fit the models for the best performance.

Instead of including a screenshot of your Kaggle submission we think that a comparative table with the tuning and performance summary of the different models built (like in Homework #5) would improve the quality of the report.

Overall/Summary

It is great that you could handle the data and implement two models, LASSO and Random Forest, successfully. With that you have laid the foundation to score high on this project. However, we think that you could put more effort in communicating your experiment and results. This includes:

- Reorganization of the document (Project understanding, Data understanding, Data preparation, Modeling and Evaluation);
- Deepen on Exploratory Data Analysis;
- Decreasing the size of images and plots, while adding more explanation;
- Be careful with the redaction, first-person singular shouldn't be used since you're a team; and
- One of the project requirements is to use related literature for the problem. This is missing at the moment.