Winning Model Documentation

A. MODEL SUMMARY

A1. Background on you/your team

- **Competition Name:** Predict Future Sales
- **Team Name:** Taras Semenchenko
- **Private Leaderboard Score:** 0.942015
- **Private Leaderboard Place:** -

- **Name:** Taras Semenchenko
- **Location:** Ukraine, Kyiv
- **Email:** semtaras20@gmail.com

A2. Background on you/your team

- **What is your academic/professional background?**

  I'm a 3rd year student.
- **Did you have any prior experience that helped you succeed in this competition?**

  Yes, I have taken various data science and machine learning courses.
- **What made you decide to enter this competition?**

  To pass the final project for "How to win a data science competition" Coursera course.
- **How much time did you spend on the competition?**

  This project took me about 4 days.

A3. Summary

For training the model I decided to choose Random Forest Regressor because it can perform high result and it is easy to tune hyperparameters. Also, this model can easily handle outliers, so they don't need to be necessary removed.

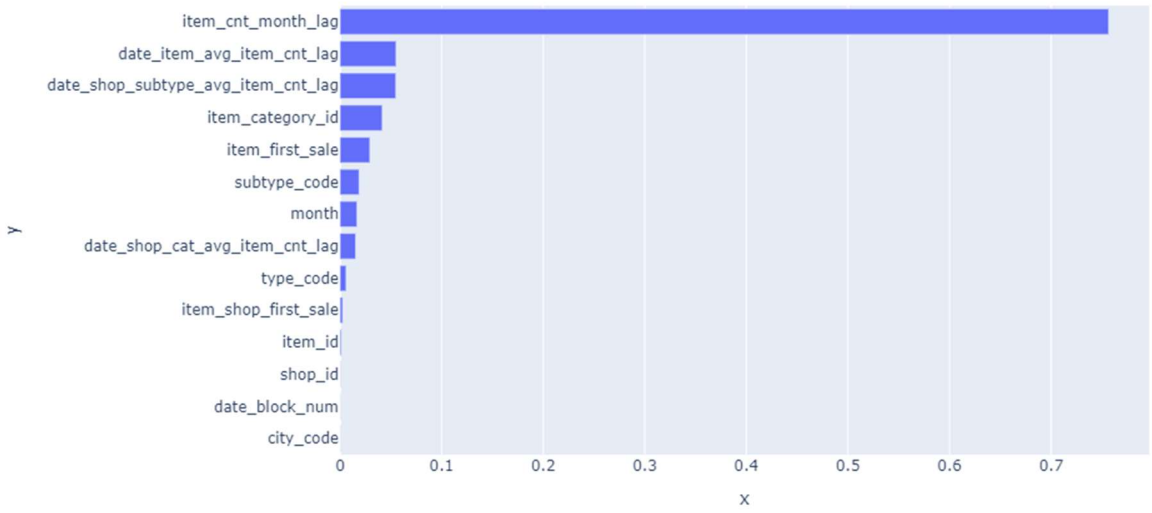
After showing feature importance I found out that feature item_cnt_month with lag 1 is the most useful (importance 76%) and this feature is number of sold items in previous month.

For training model, I took scikit-learn library because it allows to easily create and tune model.

Training my model on Kaggle kernel takes about 1554 seconds.

A4. Features Selection / Engineering

- **What were the most important features?**



- **How did you select features?**

  After using all features that was found, I chose 14 the most important ones, based on plot above.
- **Did you make any important feature transformations?**

  No, because I used tree-based model for training.

A5. Training Method(s)

- **What training methods did you use?**

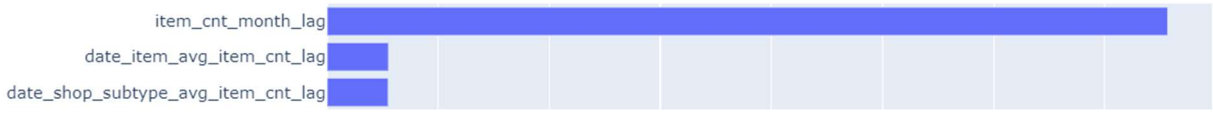
  I used only supervised training method.
- **Did you ensemble the models?**

  No, since the Random Forest model is already ensemble of many decision trees.
- **If you did ensemble, how did you weight the different models?**

  Random Forest Regressor uses the averaging method.

A7. Simple Features and Methods

- **Is there a subset of features that would get 90-95% of your final performance? Which features?**



  Yes, using the subset of features that showed above I could achieve RMSE 1.01 (on full set of features RMSE equal to 0.93).

- **What model that was most important?**

  Random Forest

- **What would the simplified model score?**

  RMSE = 1.01

A8. Model Execution Time

- **How long does it take to train your model?**

  1554 seconds

- **How long does it take to generate predictions using your model?**

  0.31 seconds for 214200 predictions

- **How long does it take to train the simplified model (referenced in section A6)?**

  574 seconds

- **How long does it take to generate predictions from the simplified model?**

  0.28 seconds for 214200 predictions

A9. References

https://www.kaggle.com/c/competitive-data-science-predict-future-sales