## horizontal line



Взгляд в будущее

10.01.2021

**─**

Tyler Arnold, Jack Bramson, Nicholas McClellan

Model Description: Our model will be trained on a Russian company’s historical sales data ranging from January 2013 to October 2015 and it will be used to predict the total sales for every product and store in November 2015.

Data Domain: We will be working with sales data provided by a Russian software company.

# Problem Statement

The problem that our group will be trying to solve is predicting the total number of sales for each product sold by 1C Company at each store selling their items in November 2015. If our model predicts the number of sales for November 2015 well given sales data from January 2013 to October 2015, 1C Company could train our model with more recent data and use it to predict future sales. There are currently no completed solutions to this problem since this is an active competition on Kaggle.

# Problem Solution

To predict the total number of sales in a month we could try different Regression models such as Ridge, Lasso or Random Forest Regression. We could also try using Support Vector Machines or Neural Networks, along with various techniques for feature selection. The model will be (primarily) evaluated on MSE, as that is the scoring criteria for the (still active) competition. To be useful, we would like the MSE to be less than 0.88263, as that would put us in the top 1,000 groups (there are over 12,000).

At this point, it’s difficult to guess how effective our model may be, but it is reasonable to expect it to be at least at the useful range, because this project was created to be the final project for a Coursera Data Science course, and this class is likely equal or greater in difficulty/amount learned. This would suggest that we should be able to perform well compared to others. Our conceptual deployment would be a web app where one could upload sales data over at least a 6 month period and receive predictions about the sales of listed items in the month following the end of the given data. However, the fallback deployment could be a simple command line interface.

# Data

We will be using the data provided by Kaggle to train our model. The datasets provided include information about each product that was sold, what day it was sold, which store sold it, etc. We shouldn’t need to go looking for more data since the data provided by Kaggle should be sufficient for training our model.

# Milestones

## Data Acquisition & Project Proposal(October 1st)

* + The data comes from the kaggle competition and has already been acquired.
  + The initial proposal for this project is also due.

## Construction of Initial Model (October 8th)

* + The group will assess which model is best to use and set up the initial (untrained) model.

## Training of Initial Model (October 15th)

* + The group will train the model and assess its performance.

## Training of Revised Model & Model Assessment (October 22nd)

* + Different model classes, or a revised model, will be implemented and analyzed to compare with the original model.

## Project Benchmark 1: Progress Video (November 1st)

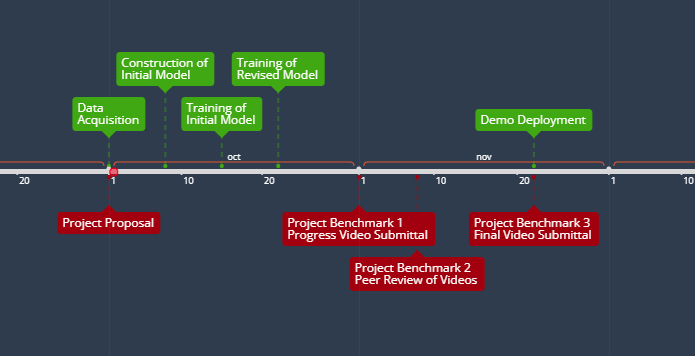
* + The second deliverable is a video showcasing the group's progress and general purpose of the project.

## Project Benchmark 2: Peer Review (November 8th)

* + The third deliverable is a peer review of another team's progress video.

## Demo Deployment & Project Benchmark 3: Final Video (November 22nd) (?)

* + A demo of the group's work will be completed in the form of a web app or command line interface.
  + The final deliverable, a video presentation, is due.



The image shown above is a timeline of the listed milestones, along with the deliverables for the course. Although the due date for the final deliverable is not known, there is a general pattern to the spacing on the timeline.

# Additional Info

Kaggle Competition & Data: https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview