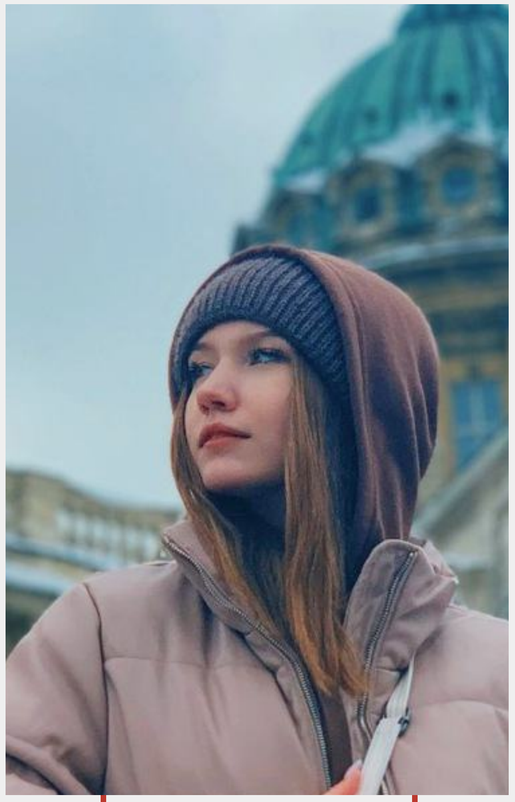
04.07.2024

АВТОРЫ: ГОЛОВИНА С.Д. БОГОМЯКОВА Г.К.
ЯКОВЛЕВА В.В. СОЛОНЬКО М.К.

КОМАНДА: ФИТ ПРЕДИКТ

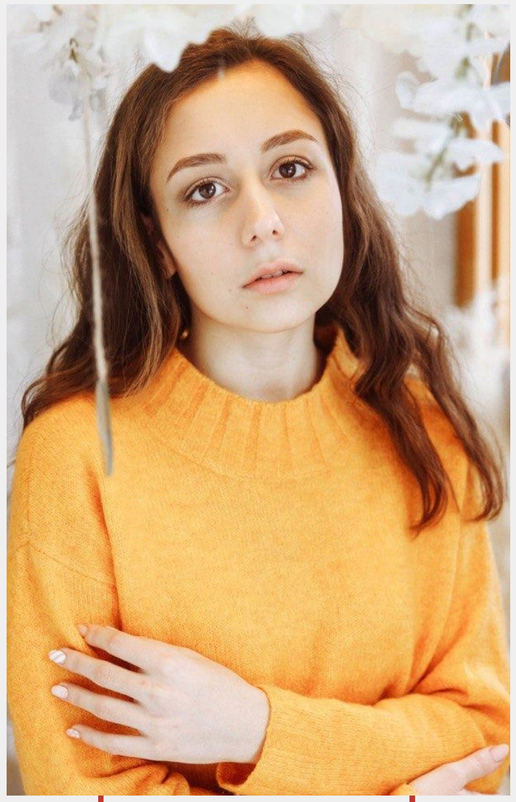
Прогнозирование сроков доставки

Наша команда



Яковлева
Валерия

Богомягкова
Екатерина



Головина
Софья

Солонько
Михаил

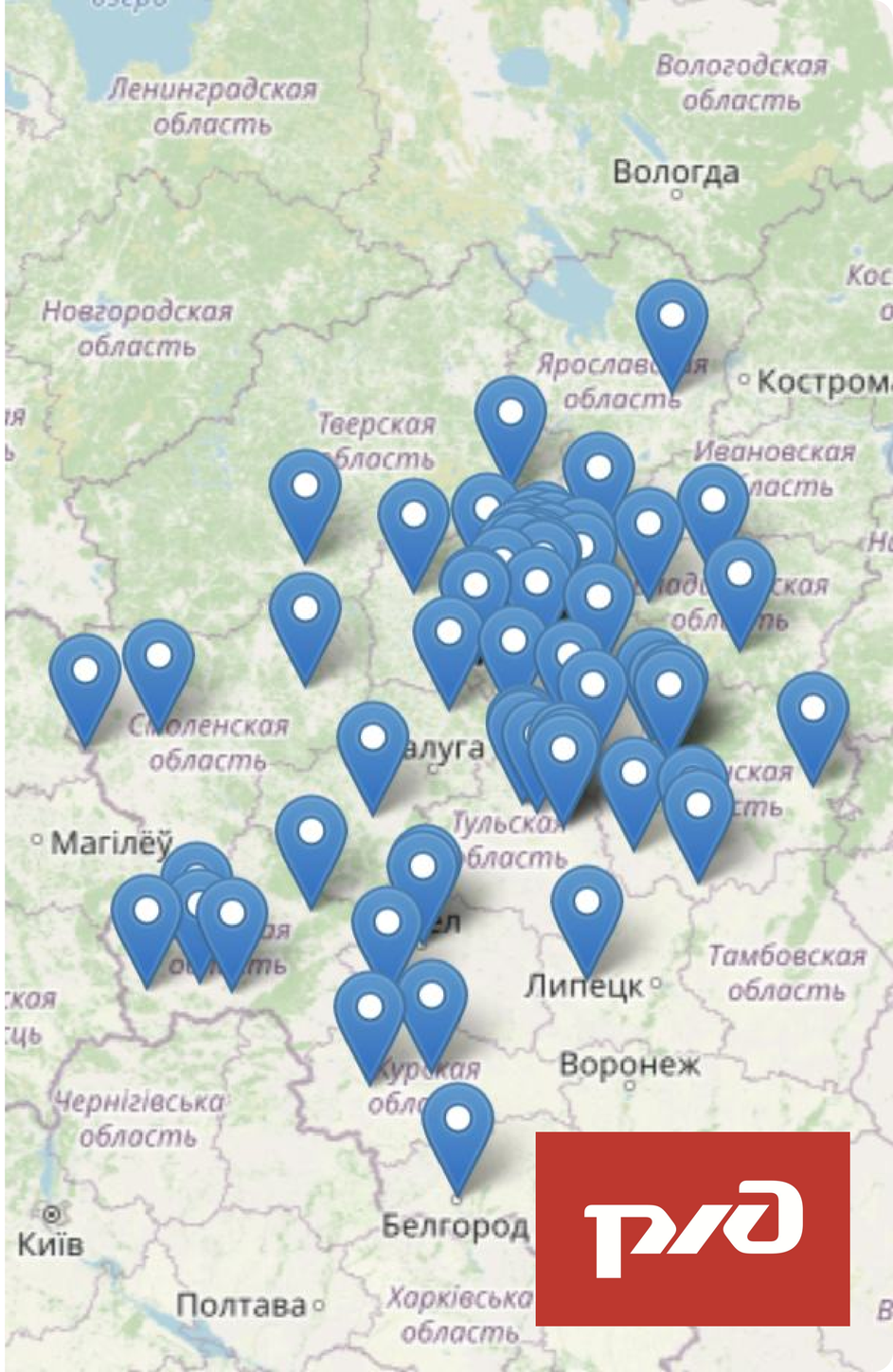
Задача

Разработать модель машинного обучения, которая будет предсказывать время доставки товаров по железной дороге из одного пункта в другой. Проверка осуществляется по MRPE.

Данные

Датасет содержит координаты исходного пункта и пункта прибытия в европейской части России, а также время, за которое поезд проезжает это расстояние в тренировочном датасете.

Также есть данные о ломанных, соединяющих пункты, и такие признаки как: направление, длина, ширина пути, а также наличие туннелей, мостов, шоссе и максимальная разрешенная скорость на участке.



Предобработка признаков

Maxspeed

Помимо пунктов отправления и прибытия будем рассматривать длину и время в пути, а также максимальную разрешенную скорость.

70% значений признака Maxspeed являются пропущенными значениями. Произвели замену пропущенных значений на 80 км/ч.

Предобработка признаков

Пропущенные пункты

Было обнаружено, что у нас отсутствуют какие-либо пути в предложенных данных у 10 станций.

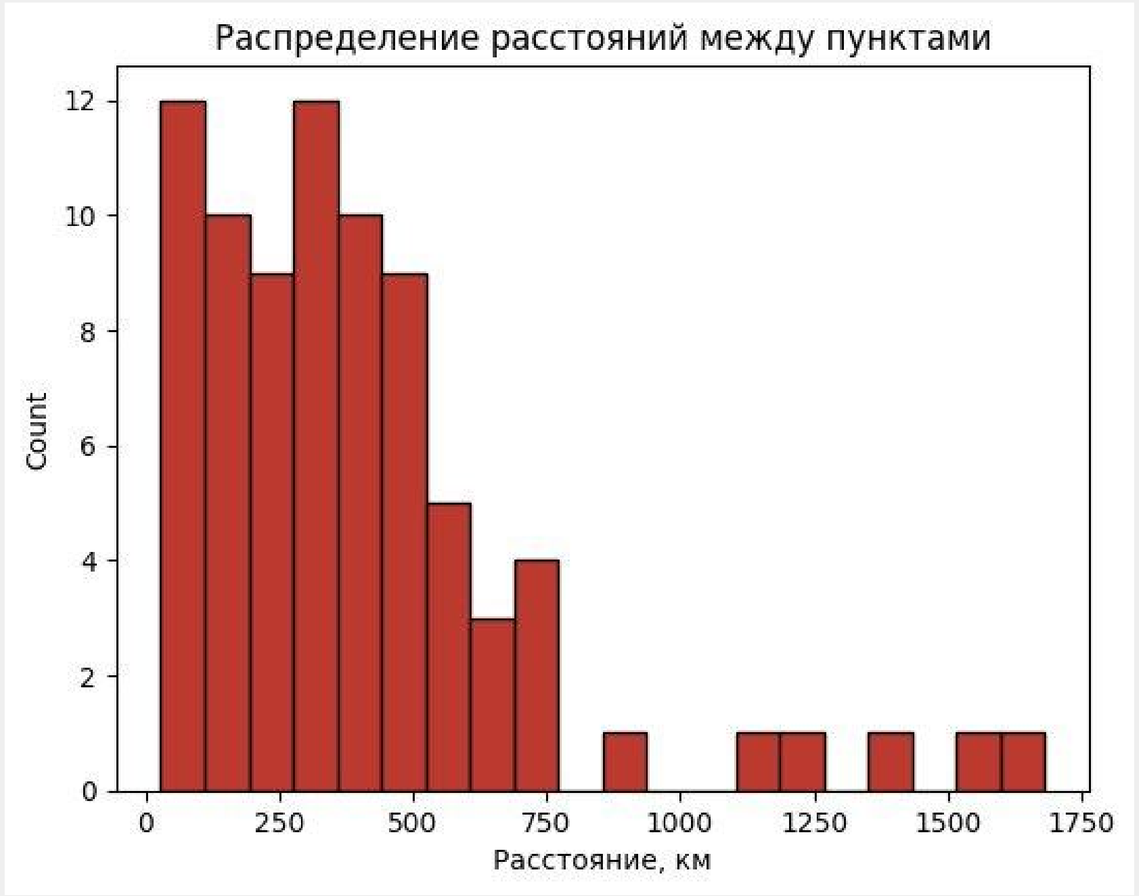
Посмотрели ближайшие ж/д станции к данным пунктам и нашли приблизительные расстояния по ж/д путям между ними и добавили новые длины путей.

Аэропорт
Бекасово-Центральное
Брянск-Орловский
Домодедово
Люберцы II
Павелец-Тульский
Перово
Степенькино-II
Тула-Вяземская
Узловая I

Аггрегация

Исходные данные - это точки маршрута с определенными характеристиками.

Соединим их по пункту отправления и прибытия, представив параметр просуммировав length и представив его в км, а maxspeed как среднее арифметическое скоростей.

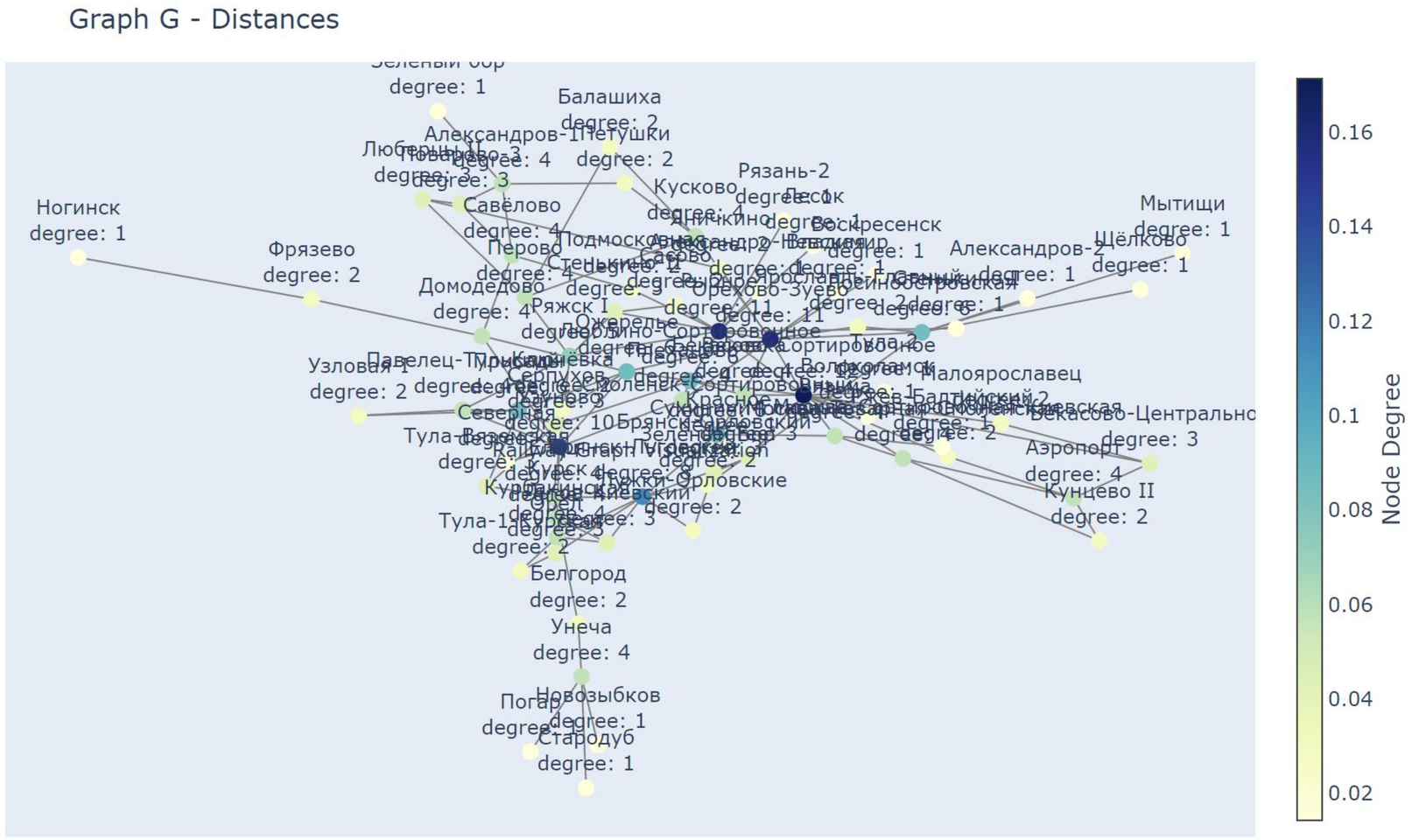


Граф по длине пути

Используем поиск кратчайшего пути между двумя станциями (узлами) в графе G.

Он проверяет наличие станций в графе, затем использует алгоритмы NetworkX для нахождения кратчайшего пути и его длины.

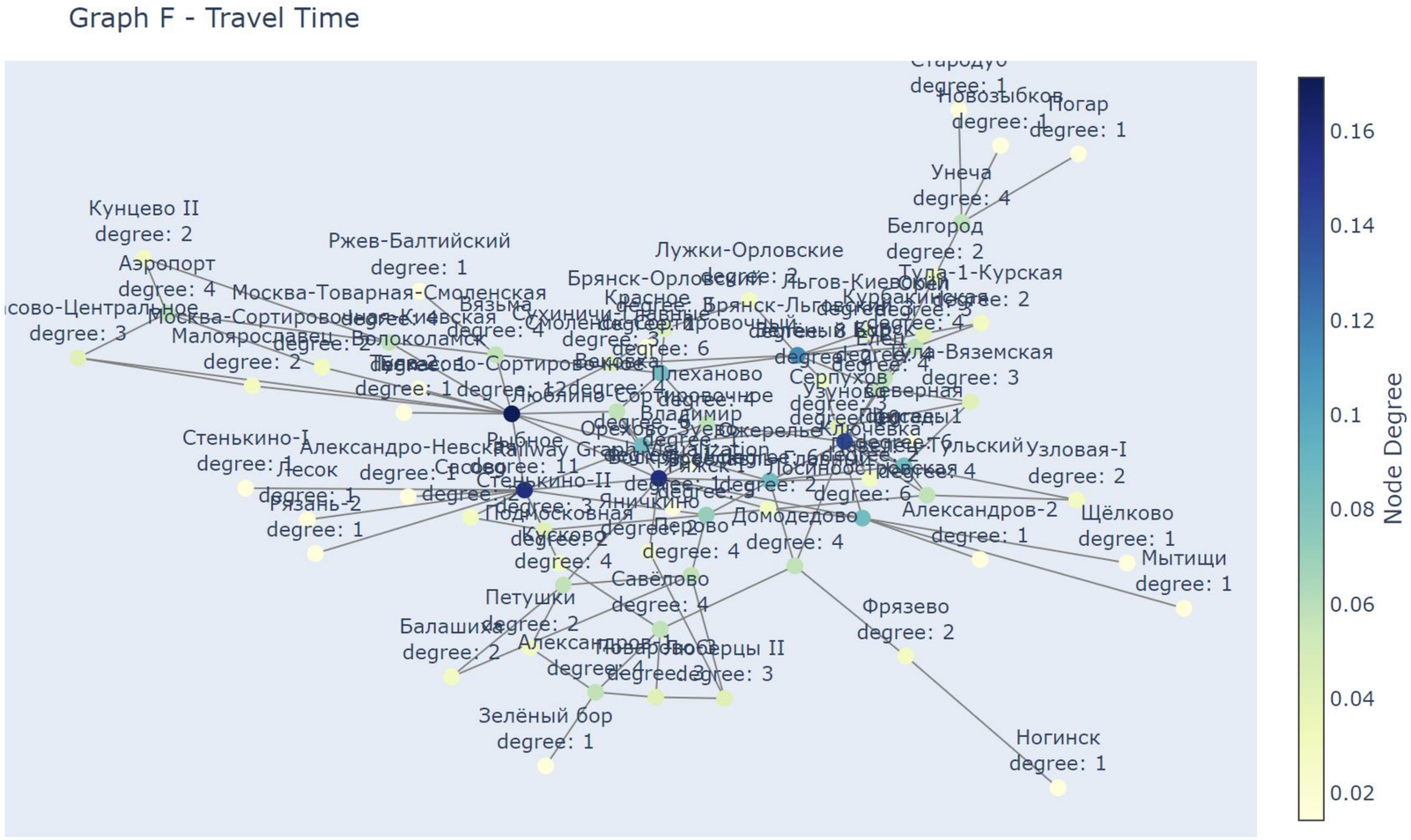
Дополнительно считаем количество станций на пути, а также минимальное и максимальное расстояние между станциями.



Граф по времени в пути

Работает с графом, у которого веса ребер - минимальное время, за которое можно проехать между станциями, если развить максимальную скорость, и ищет самый быстрый путь между станциями.

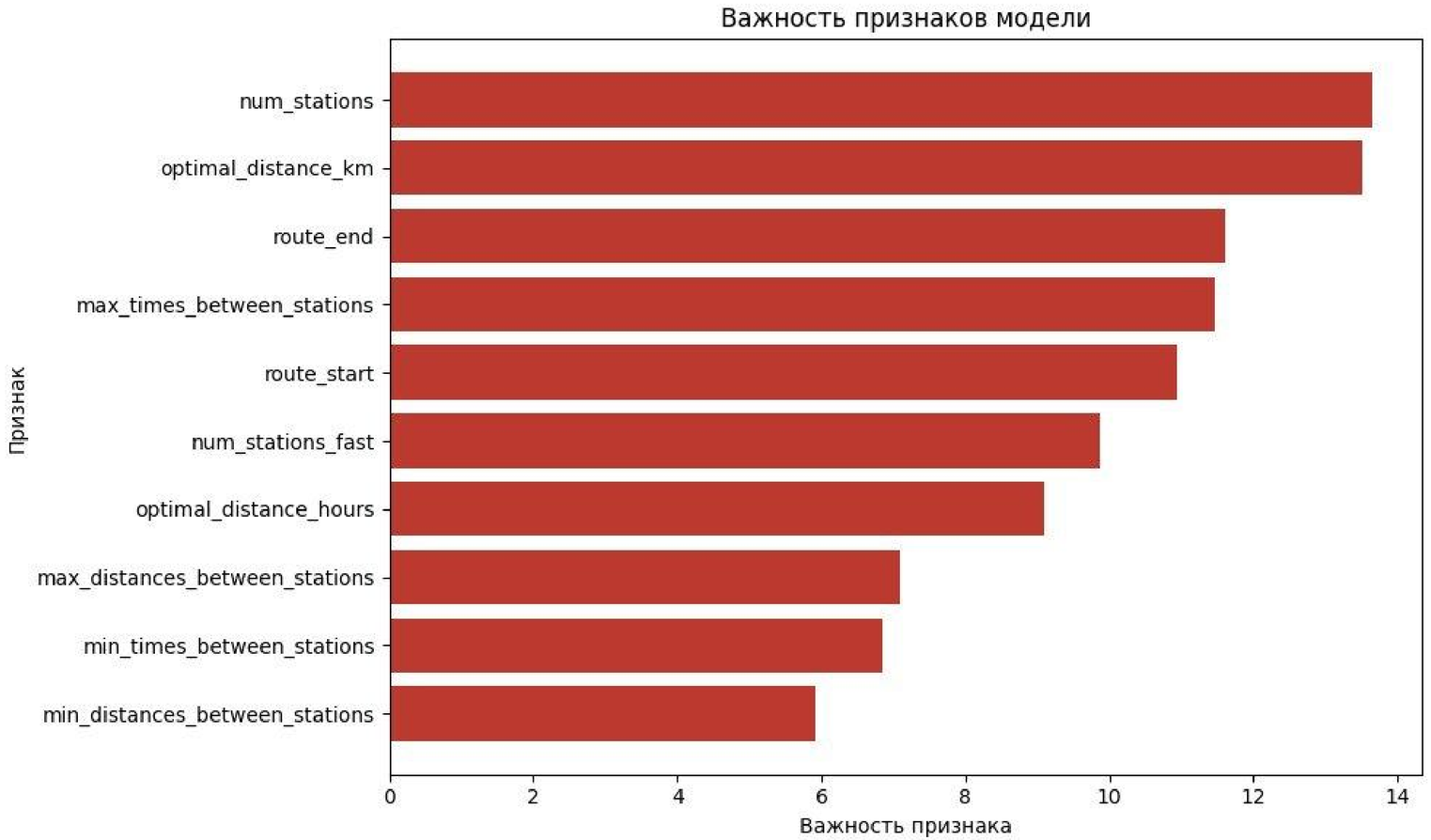
Он использует алгоритм Дейкстры для поиска пути и его длительности, а также вычисляет количество станций и минимальное и максимальное время между соседними станциями на найденном пути.



CatBoost

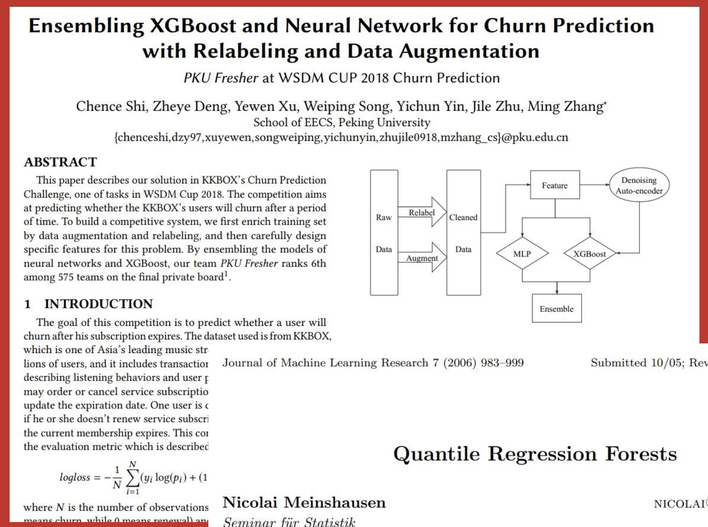
Гиперпараметры:

- learning rate = 0.1
- depth = 6
- l2_leaf_reg = 3
- loss_function = ‘RMSE’
- iterations = 1499



Прогнозы, прошедшие тест: 1503 из 1519
MRPE: 8.06%

Тестируемые модели



Ensembling XGBoost and Neural Network for Churn Prediction with Relabeling and Data Augmentation

PKU Fresher at WSDM CUP 2018 Churn Prediction

Chence Shi, Zheye Deng, Yewen Xu, Weiping Song, Yichun Yin, Jile Zhu, Ming Zhang*
School of EECS, Peking University
{chenceshi,dzy97,xuyewen,songweiping,yichunyin,zhujile0918,mzhang_cs}@pku.edu.cn

ABSTRACT

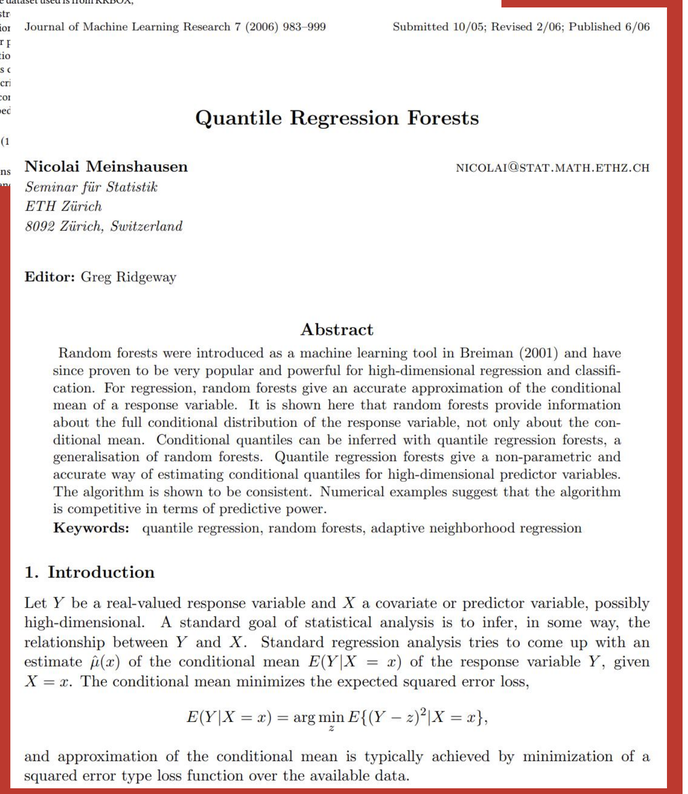
This paper describes our solution in KKBOX's Churn Prediction Challenge, one of tasks in WSDM Cup 2018. The competition aims at predicting whether the KKBOX's users will churn after a period of time. To build a competitive system, we first enrich training set by data augmentation and relabeling, and then carefully design specific features for this problem. By ensembling the models of neural networks and XGBoost, our team *PKU Fresher* ranks 6th among 575 teams on the final private board[1].

1 INTRODUCTION

The goal of this competition is to predict whether a user will churn after his subscription expires. The dataset used is from KKBOX, which is one of Asia's leading music str
lions of users, and it includes transactio
describing listening behaviors and user
may order or cancel service subscriptio
update the expiration date. One user is
if he or she doesn't renew service subscr
the current membership expires. This co
the evaluation metric which is describe

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1$$

where N is the number of observations
means churn, while 0 means renewal suc



Journal of Machine Learning Research 7 (2006) 983–999 Submitted 10/05; Revised 2/06; Published 6/06

Quantile Regression Forests

Nicolai Meinshausen NICOLAI@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
8092 Zürich, Switzerland

Editor: Greg Ridgeway

Abstract

Random forests were introduced as a machine learning tool in Breiman (2001) and have since proven to be very popular and powerful for high-dimensional regression and classification. For regression, random forests give an accurate approximation of the conditional mean of a response variable. It is shown here that random forests provide information about the full conditional distribution of the response variable, not only about the conditional mean. Conditional quantiles can be inferred with quantile regression forests, a generalisation of random forests. Quantile regression forests give a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables. The algorithm is shown to be consistent. Numerical examples suggest that the algorithm is competitive in terms of predictive power.

Keywords: quantile regression, random forests, adaptive neighborhood regression

1. Introduction

Let Y be a real-valued response variable and X a covariate or predictor variable, possibly high-dimensional. A standard goal of statistical analysis is to infer, in some way, the relationship between Y and X. Standard regression analysis tries to come up with an estimate $\hat{\mu}(x)$ of the conditional mean $E(Y|X = x)$ of the response variable Y, given $X = x$. The conditional mean minimizes the expected squared error loss,

$$E(Y|X = x) = \arg\min_z E\{(Y - z)^2|X = x\},$$

and approximation of the conditional mean is typically achieved by minimization of a squared error type loss function over the available data.

Quantile RF

Квантильный регресионный лес. Предсказывает 5 и 95 перцентиль времени доставки. Сместили баланс на 95 перцентиль, ошибка 42%. Можно было поиграть с балансом и получить ошибку меньше. Данные выбросов не имеют, поэтому квантильные модели не лучшее решение

FCNN

Полносвязные нейронные сети с регуляризацией. Архитектура была примитивной, сама идея возможно не лучшая для этой задачи, много затрат на предобработку данных. Ошибка 23%

Ensemble 0,6CatBoost + 0,4FCNN

Ансамбль из CatBoost и полносвязной нейронной сети. Идея не оправдала ожиданий, скорее всего из-за плохой НН, ошибка 34%

GNN

Графовая нейронная сеть. Интересно, но незнакомо.

Инсайты

Максимальная скорость

Лучшие метрики получаются при использовании именно 80 км/ч, а не стандартных 90 км/ч

Использование средней гармонической для скорости также дает показатели хуже, чем для средней арифметической.

Путь

Отсутствуют вложенные отрезки, а также получившаяся из датасета длина пути чуть больше, чем длина пути, построенная в Яндекс.Картах.

Спасибо за внимание!

Фит предикт