Разработка нейросетевой модели для анализа спутниковых данных для оценки урожайности сельскохозяйственных культур в районах РФ

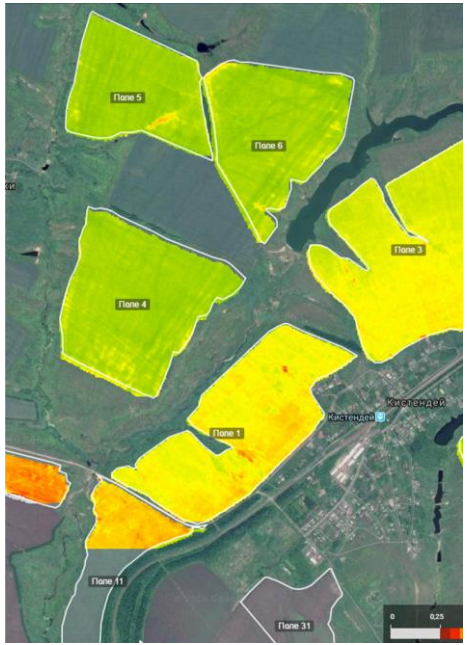
Сычков Александр Александровчи

Научный консультанта на курсе MSU.AI - Ганичев Антон Александрович

Научный руководитель - Трошко Ксения Анатольевна

Введение в тему

Задача состоит в том, чтобы имея в распоряжении спутниковые снимки и данные о погоде иметь возможность оценить текущее состояние посевов, выраженное через расчетную урожайность, в любое время сезона. С помощью спутниковых снимков расчитывают т.н. Normalized Difference Vegetation Index. Это спектральный индекс, коррелирующий с густотой растительности и с содержанием хлорофила. Традиционно используется для оценки состояния растений по мультиспектральным снимкам.



С помощью метеорологических спутников и системы реанализа, можно получать сведения о погодных условиях в любой точке мира. Задача состоит в том, чтобы научиться анализировать эти данные и давать заблаговременный прогноз.

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

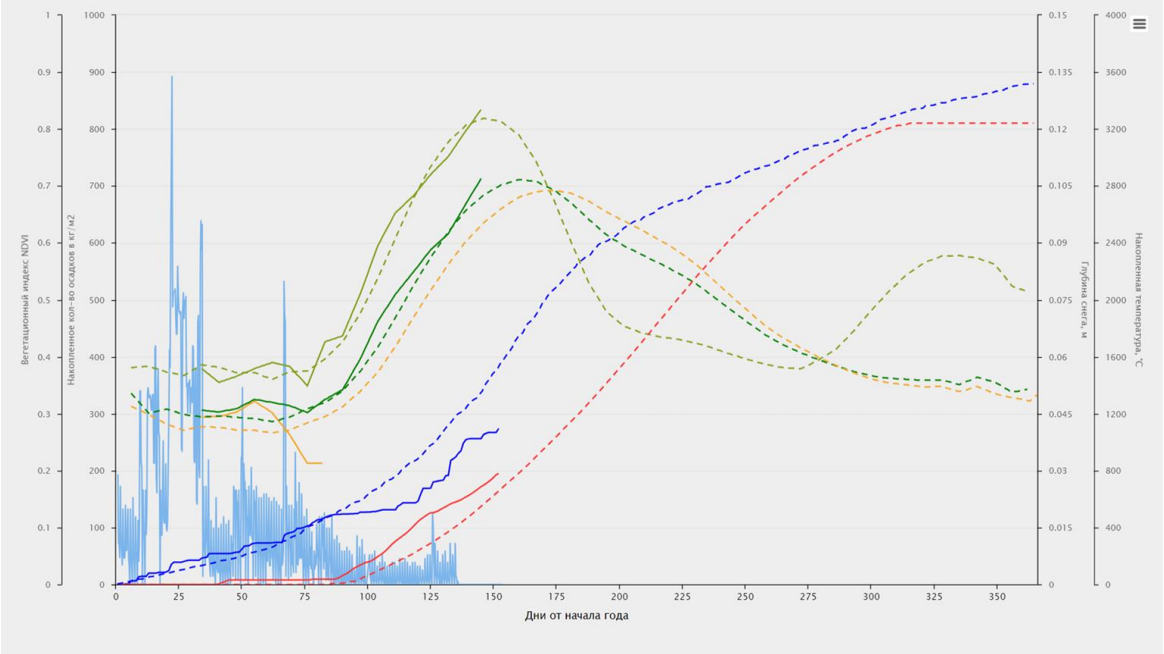
Данные

Данные, которые необходимо научиться обрабатывать, представляют собой временные ряды спектрального индекса NDVI и набора погодных параметров, доступных в системе реанализа National Centers for Environmental Prediction. Целевые значения - значения урожайности озимой пшеницы в районах России, полученные из Базы Данных Муниципальных Образований. Удалось получить информацию для 797 административных районов в период с 2007 по 2022 год.

Источник временных рядов NDVI – ВЕГА-Science. Временные ряды формируются на конкретных территориях, детектированных как маски земель, занятых озимыми культурами, на основе еженедельных композитов спектрометра MODIS. (Спутники Terra и Aqua)

Важным нюансом является то, что временные ряды имеют разную дискретность и требуют приведение к единому стандарту размерности путем усреднения или интерполяции.

Данные



Примеры временных рядов имеющихся данных

Модели

- ●Главным образом рассматривается модель с рекурентным энкодером с модулем внимания, который будет создавать эмбединг и передавать его в полносвязную сеть для анализа. Однако в дальнейшем был сделан выбор в пользу применения сверточных слоев.
- ●В первом приближении предпринимались попытки обучить полносвязную сеть на загрубленных даных и использовать сверточные слои вместо энкодера.
- ●На данный момент получено несколько моделей: Случайный лес, полносвязная сеть, сеть со сверточными слоями в качестве энкодера.
- ●Так как речь идет об обычной задаче регрессии, в качестве loss функции использовалось среднеквадратичное отклонение MSE. В качестве дополнительных метрик качества использовались MAE, RMSE и R^2

Обзор существующих решений

Модель	MSE	MAE	RMSE	R^2
Средняя урожайность в районе как прогноз	69.09	6.46	8.31	0.55
Линейно регрессионная модель с индивидуальным построением для регионов, примененная ко всей тестовой выборке.	63.06	6.41	7.94	0.27
Линейно регрессионная модель с индивидуальным построением для регионов, примененная только к регионам с высокой корреляцией ndvi_max / productive	42.33	5.22	6.39	0.54

Модели классического ml и их метрики для озимой пшеницы.

Модель	Комментарий	MSE	MAE	RMSE	R^2
Линейная модель	-	77.324	6.988	8.793	0.509
Случайное дерево	Максимальная глубина: 30	45.866	5.179	6.772	0.696
Случайный лес	Глубина: 50 Деревьев: 200	22.436	3.61	4.737	0.851
Случайный лес	Глубина: 100 Деревьев: 200	**22.161**	**3.603**	**4.708**	**0.853**

Предварительные опыты по использованию классических методов машинного обучения для озимой пшеницы .

Датасет v1

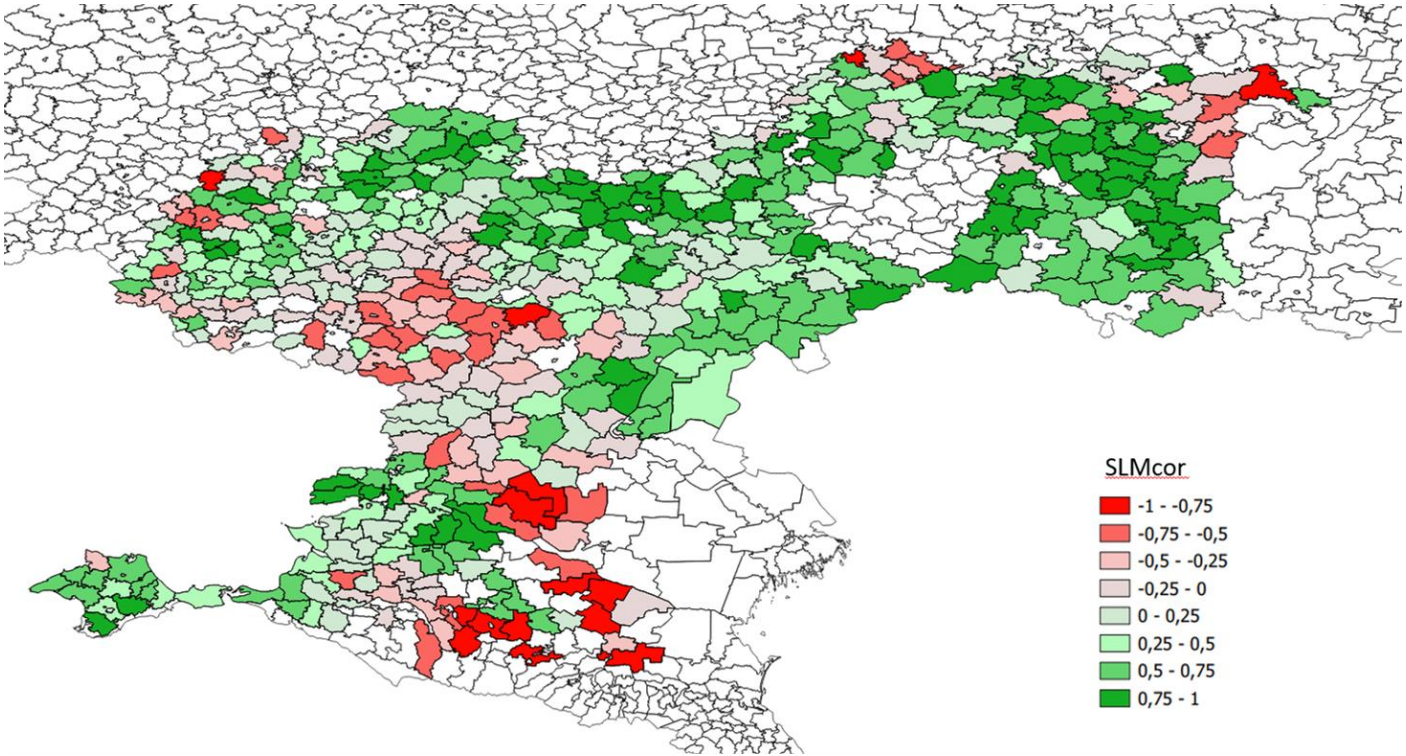
Модель	Комментарий	MSE	MAE
Полносвязная нейронная сеть	Кол-во слоев: 8. Функция активации: Softplus. Эпох обучения: 200. Конфигурация: 500 нейронов в каждом скрытом слое	**26.61**	**4.03**
Полносвязная нейронная сеть	Кол-во слоев: 15. Функция активации: Softplus. Эпох обучения: 200. Конфигурация: Первый скрытый слой имеет 200 нейронов и их количество постепенно снижается до 10	30.34	4.27
Полносвязная нейронная сеть	Кол-во слоев: 15. Функция активации: ReLu. Эпох обучения: 200 Конфигурация: Первый скрытый слой имеет 200 нейронов и их количество постепенно снижается до 10	29.32	4.12

Изменения в датасете

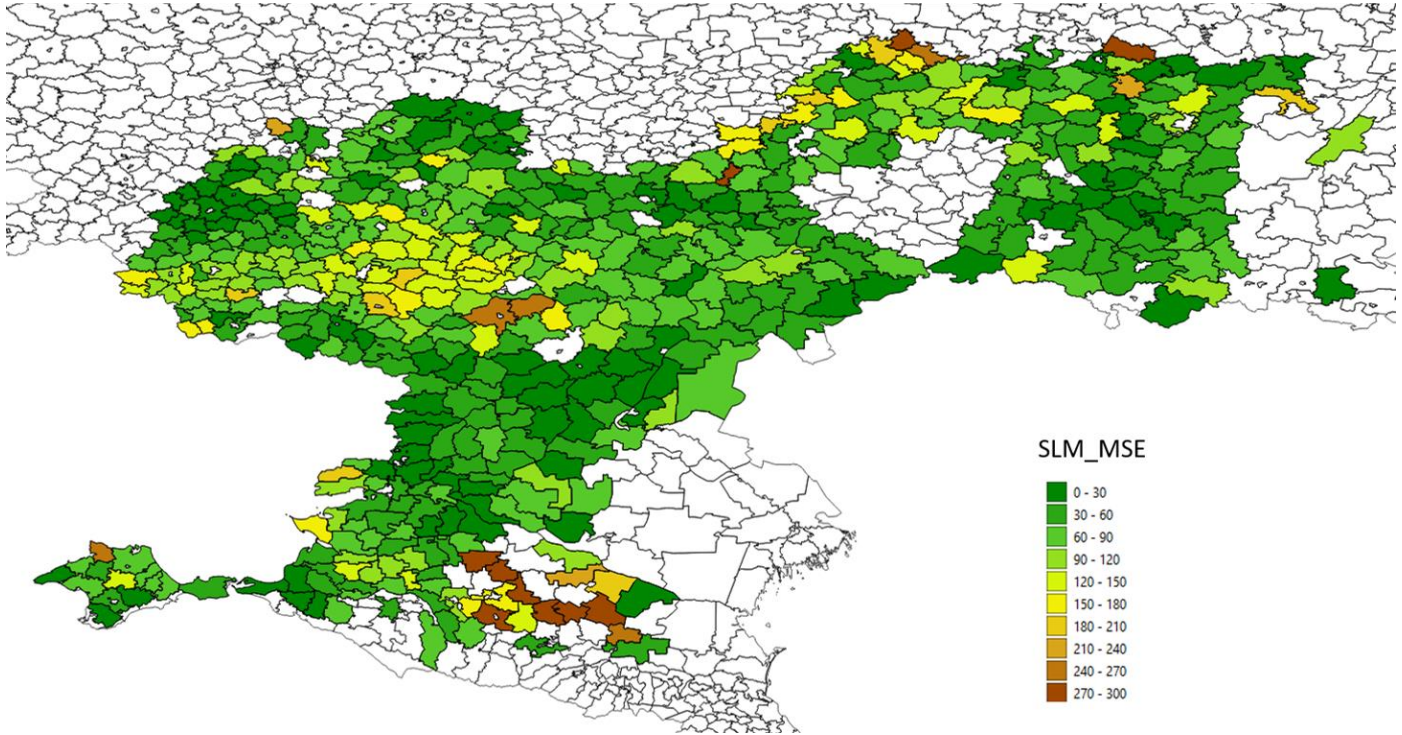
В первом датасете одним из входных параметров была средняя урожайность в районе в прошлые годы. Она была заменена на скользящую среднюю урожайность за 5 последних лет. Кроме того, в список входных данных был добавлен скользящий тренд изменения урожайности.

В дальнейшем была написана единая библиотека, которая используется для подготовки данных для обучения и для непосредственных расчетов будущей урожайности. Это важный момент, так как использование одной библиотеки позволяет снизить риск отличий в процедуре подготовки данных. Данные для предсказания и данные для обучения должны быть подготовлены идентичным образом.

Корреляция при традиционном методе



MSE при традиционном методе



Результаты нового опыта

Модель	MAE	MSE	RMSE	R^2
Нейронная сеть (Тестовая выборка до изменений в датасете)	3,43	20,54	4,53	-
Нейронная сеть (Тестирование на ретроспективных данных до изменений в алгоритме подготовки данных)	**5,477**	**49,924**	**7,06**	**0,413**
Нейронная сеть (Тестирование на ретроспективных данных после улучшений в коде подготовке данных)	**3.629**	**23.436**	**11.781**	**-**
Линейная регрессия	6,731	75,120	8,66	0,154

Модели классического ml и их метрики для озимой пшеницы

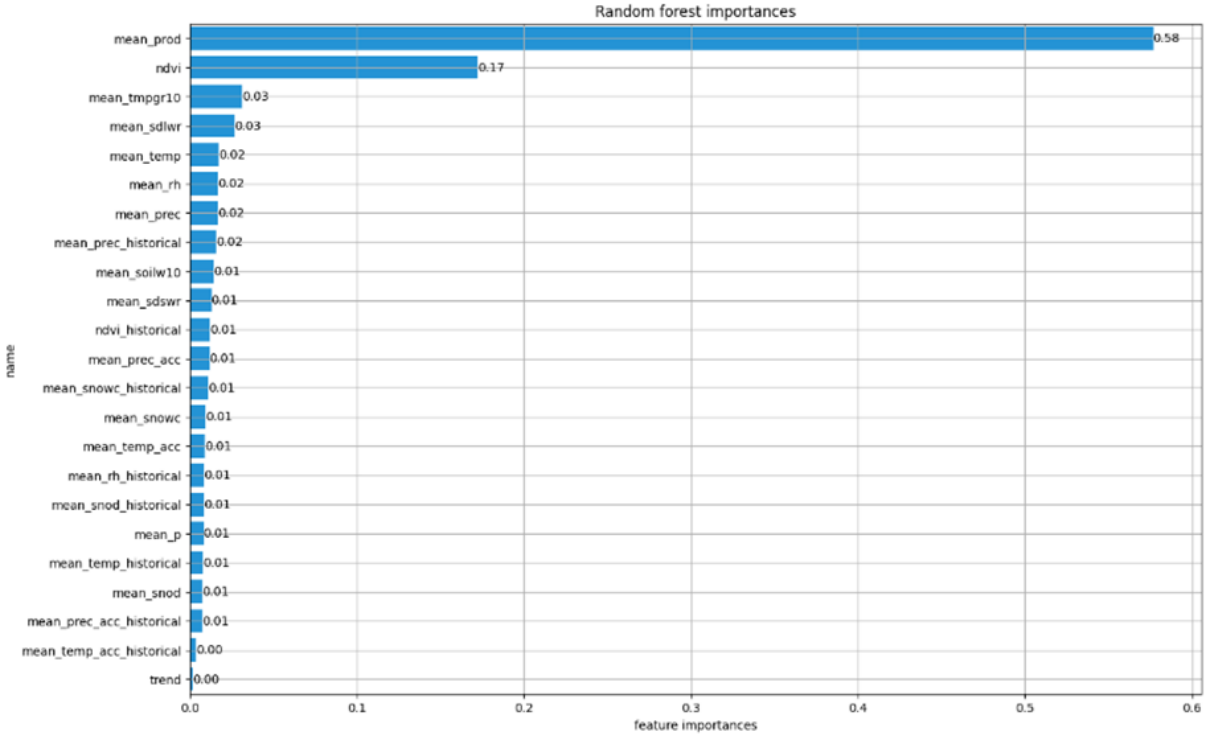
Модель	Комментарий	MSE	MAE	RMSE	R^2
Линейная модель	-	25.74	3.863	5.074	0.866
Случайное дерево	Максимальная глубина: 50	53.93	5.511	7.344	0.719
Случайный лес	Глубина: 70 Деревьев: 25	25.144	3.825	5.014	0.869
Случайный лес	Глубина: 50 Деревьев: 200	**24.195**	**3.727**	**4.919**	**0.874**

Предварительные опыты по использованию классических методов машинного обучения после изменений в датасете

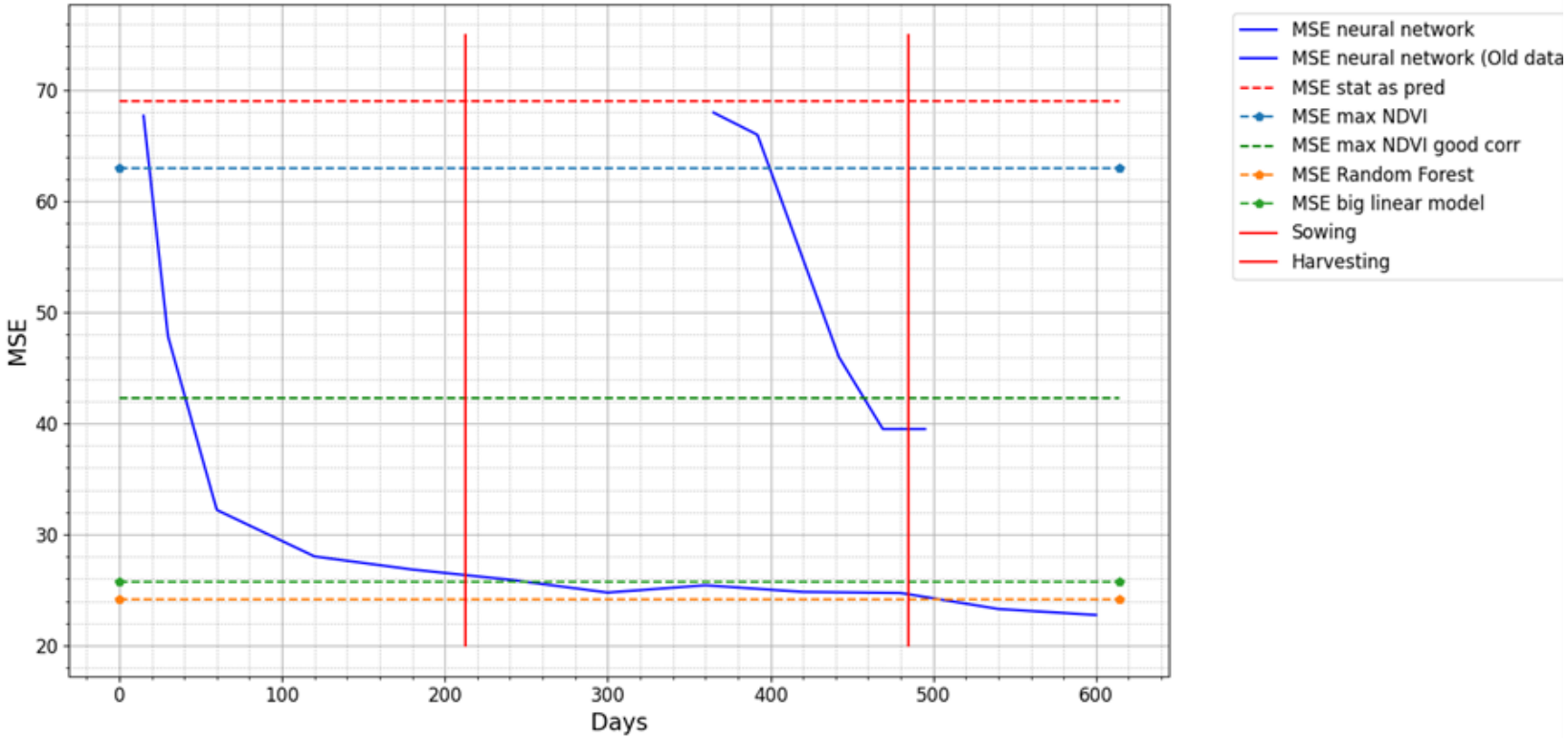
Оценка важности параметров для озимой пшеницы

№	Missing feature	Лучший MSE	MAE	MSE пяти лучших моделей	MAE пяти лучших моделей
1	ndvi / ndvi_historical /	26,68	3,96	26,98	4,00
2	ndvi /	26,01	3,85	26,43	3,88
3	ndvi_historical /	23,73	3,65	23,98	3,67
4	Используются все даннные	23,66	3,60	24,15	3,64
5	mean_temp /	23,94	3,60	24,12	3,65
6	mean_temp_historical /	23,47	3,69	24,18	3,73
7	mean_temp_acc /	24,01	3,64	24,22	3,71
8	mean_temp_acc /	24,17	3,63	24,25	3,67
9	mean_temp_acc_historical /	23,53	3,67	24,16	3,70
10	mean_prec /	24,02	3,67	24,22	3,71
11	mean_prec_historical /	23,76	3,63	24,13	3,67
12	mean_prec_acc /	24,00	3,62	24,37	3,66
13	mean_prec_acc_historical /	23,59	3,64	23,97	3,66
14	mean_rh /	23,89	3,64	24,19	3,68
15	mean_rh_historical /	23,60	3,64	23,90	3,68
16	mean_p /	24,51	3,66	24,82	3,68
17	mean_snod /	23,39	3,60	23,56	3,63
18	mean_snod_historical /	23,41	3,61	23,96	3,64
19	mean_snowc /	24,04	3,65	24,35	3,67
20	mean_snowc_historical /	23,82	3,61	24,23	3,64
21	mean_sdswr /	23,60	3,60	23,87	3,61
22	mean_sdlwr /	22,64	3,52	23,07	3,58
23	mean_tmpgr10 /	23,57	3,58	23,76	3,62
24	mean_soilw10 /	23,77	3,67	24,28	3,70
25	mean_prod /	29,87	4,13	30,95	4,21
26	trend /	22,75	3,55	23,77	3,65
27	disp /	22,76	3,61	23,11	3,64
28	Без исторических данных и продуктивности	35,54	4,51	37,43	4,66
29	Без исторических и интегральных данных	22,66	3,59	22,89	3,64

Информативность параметров для сои согласно случайному лесу



Возможности заблаговременного прогноза для озимой пшеницы



Модели классического ml и их метрики для сои

Модель	Комментарий	MSE	MAE	RMSE	R^2
Линейная модель	-	**10.347**	**2.496**	**3.217**	**0.685**
Случайное дерево	Максимальная глубина: 50	22.285	3.666	4.721	0.322
Случайный лес	Глубина: 70 Деревьев: 25	11.146	2.58	3.339	0.661
Случайный лес	Глубина: 200 Деревьев: 200	**10.941**	**2.592**	**3.308**	**0.667**

Предварительные опыты по использованию классических методов машинного обучения для сои

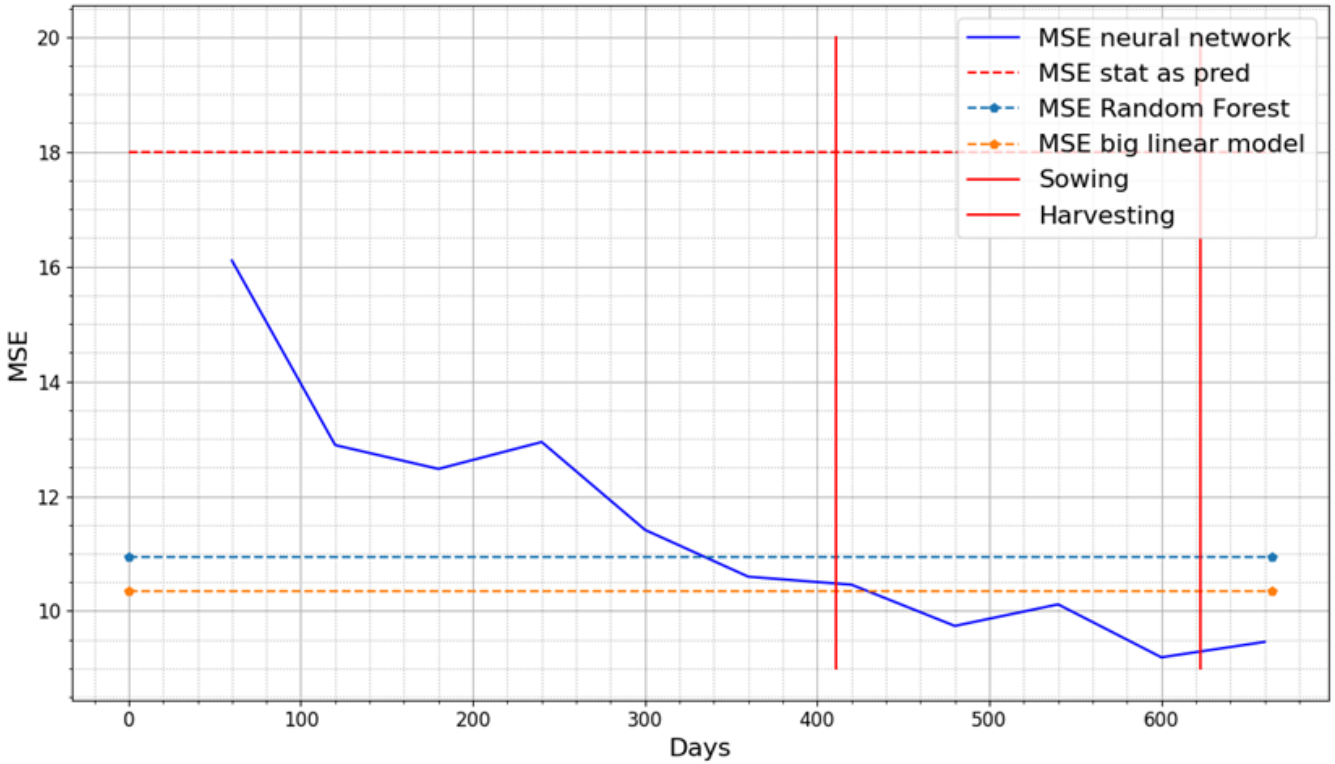
Метрики нейронных сетей для сои

Комментарий	Loss train MSE	Loss test MSE	Test MSE
Нейросетевая модель, принимающая на вход только среднюю продуктивность. Датасет не подвергался очистке.	18.10	24.70	3.26
Нейросеть, принимающая на вход среднюю скользящую урожайность, дисперсию и тренд. Датасет не подвергался очистке.	24.65	25.76	3.88
Нейросеть, принимающая на вход все доступные данные. Датасет не подвергался очистке.	**13.31**	**17.24**	**3.07**

Оценка значимости параметров для сои

№	Исключенные параметры	MSE лучшей модели	MAE лучшей модели	MSE 5 лучших моделей	MAE 5 лучших моделей
1		9,51	2,64	9,90	2,69
2	ndvi /	10,65	2,76	10,96	2,77
3	ndvi / ndvi_historical /	10,86	2,74	11,11	2,76
4	ndvi_historical /	9,25	2,70	9,65	2,71
5	mean_temp /	9,53	2,65	10,17	2,70
6	mean_temp_historical /	9,51	2,70	10,02	2,73
7	mean_temp_acc /	9,77	2,66	9,87	2,67
8	mean_temp_acc_historical /	9,32	2,61	9,52	2,64
9	mean_prec /	10,82	2,69	11,26	2,76
10	mean_prec_historical /	10,03	2,66	10,20	2,70
11	mean_prec_acc /	8,91	2,64	9,22	2,66
12	mean_prec_acc_historical /	9,17	2,61	9,24	2,63
13	mean_rh /	9,73	2,73	9,93	2,76
14	mean_rh_historical /	9,10	2,63	9,46	2,66
15	mean_p /	9,20	2,66	9,45	2,68
16	mean_p_historical /	9,58	2,65	9,68	2,68
17	mean_snod /	8,75	2,66	9,22	2,68
18	mean_snod_historical /	9,45	2,60	9,67	2,66
19	mean_snowc /	8,95	2,61	9,22	2,63
20	mean_snowc_historical /	8,81	2,65	9,50	2,70
21	mean_htc_decade /	9,51	2,62	9,80	2,66
22	mean_sdswr /	9,22	2,61	9,64	2,63
23	mean_sdswr_historical /	9,48	2,62	9,70	2,65
24	mean_sdlwr /	9,08	2,64	9,41	2,70
25	mean_sdlwr_historical /	9,25	2,67	9,51	2,69
26	mean_tmpgr10 /	9,12	2,65	9,26	2,67
27	mean_soilw10_historical /	9,93	2,65	10,20	2,70
28	mean_prod /	20,30	4,06	20,69	4,09
29	trend /	9,53	2,72	9,90	2,72
30	disp /	9,23	2,66	9,89	2,72
31	Все исторические параметры	8,78	2,52	9,44	2,55
32	Все исторические и интегральные параметры	8,31	2,48	8,54	2,51

Возможности заблаговременного прогноза для сои



Модели классического ml и их метрики для яровой пшеницы

Модель	Комментарий	MSE	MAE	RMSE	R^2
Случайный лес	Число деревьев:25 Максимальная глубина: 70	27.435	**3.787**	**5.238**	0.793
Случайный лес	Число деревьев:50 Максимальная глубина: 200	**26.361**	**3.721**	**5.134**	**0.801**
Решающее дерево	Максимальная глубина: 50	53.304	5.381	7.301	0.598
Решающее дерево	Максимальная глубина: 200	55.051	5.518	7.420	0.585
Линейная модель		31.82	4.258	5.641	0.760

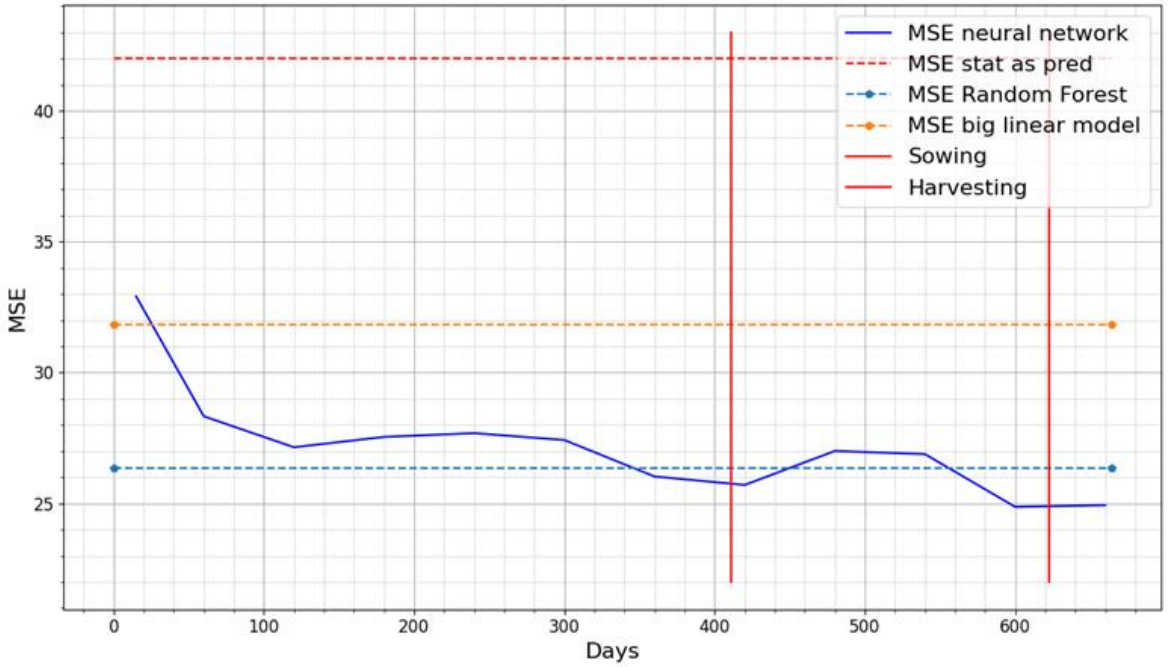
Метрики нейронных сетей для яровой пшеницы

Комментарий	Test RMSE	Loss test MSE	Test MAE
Нейросетевая модель, принимающая на вход только среднюю продуктивность.	7.01	49,04	5,28
Нейросеть, принимающая на вход все доступные данные	5.36	28,35	3,91

Оценка значимости параметров для яровой пшеницы

№	Исключенные параметры	MSE лучшей модели	MAE лучшей модели	MSE 5 лучших моделей	MAE 5 лучших моделей
1		26,65	4	27,35	4,04
2	ndvi /	29,19	4,09	29,68	4,16
3	ndvi / ndvi_historical /	28,38	4,19	28,9	4,24
4	ndvi_historical /	26,82	3,96	27,3	4,01
5	mean_temp /	25,35	3,97	26,78	4,02
6	mean_temp_historical /	25,21	3,93	26,74	3,98
7	mean_temp_acc /	26,46	3,97	27,38	4,05
8	mean_temp_acc_historical /	26,11	4	27,33	4,07
9	mean_prec /	26,62	3,94	27,62	4,01
10	mean_prec_historical /	25,31	3,92	26,55	3,98
11	mean_prec_acc /	27	4	27,38	4,03
12	mean_prec_acc_historical /	27,52	4,05	28,19	4,17
13	mean_rh /	28,01	4,06	28,81	4,16
14	mean_rh_historical /	26,45	3,88	27,46	3,97
15	mean_p /	27,11	3,93	27,78	4,03
16	mean_p_historical /	27	4	27,95	4,05
17	mean_snod /	27,22	4,09	28,56	4,13
18	mean_snod_historical /	28,84	4,14	29,11	4,18
19	mean_snowc /	26,19	4	27,33	4,1
20	mean_snowc_historical /	27,46	4,1	28,92	4,21
21	mean_htc_decade /	27,35	4,09	28,87	4,15
22	mean_htc_decade_historical /	25,15	3,79	25,85	3,86
23	mean_sdswr /	26,7	3,88	27,63	3,99
23	mean_sdswr_historical /	27,59	3,98	28,38	4,05
24	mean_sdlwr /	25,26	3,93	27,09	4,02
25	mean_sdlwr_historical /	27,05	4,02	27,85	4,08
26	mean_tmpgr10 /	25,01	3,85	26,21	3,93
27	mean_soilw10_historical /	27,76	3,97	28,81	4,09
28	mean_prod /	84,06	6,98	84,56	7,01
29	trend /	26,5	3,97	26,94	4,01
30	disp /	27,64	3,98	29,17	4,07
31	Все исторические параметры и дисперсия	26,59	3,97	27,02	4
32	Все исторические, интегральные параметры и дисперсия.	25,88	3,96	26,57	4
33	Все параметры кроме средней продуктивности.	32,73	4,63	32,83	4,63

Возможности заблаговременного прогноза для яровой пшеницы



Резюме опытов

Культура	Скользящее среднее как прогноз [MSE]	Линейная модель [MSE]	Случайный лес [MSE]	Нейронная сеть [MSE]	Прирост точности в [%]
Озимая пшеница	69.83	25.74	24.195	**22.66**	67.54
Яровая пшеница	52.06	31.82	**18.985**	19.06	63.38
Соя	18.10	13.298	10.0491	**8.54**	52.81

Сравнение с другой результатом другой статьи

Модель	Комментарий	MSE [ц/га]^2	RMSE [ц/га]
CNN-RNN	Модель из исследования A CNN-RNN Framework for Crop Yield Prediction	9.59	3.097
CNN (our)	Модель данного исследования	8.54	2.922

План

1) Внедрить модель. Написать программу для сбора данных и составления отчетов. ✓
2) Применить ✓
3) По результатам опубликовать статью
4) Произвести опыты с прогнозированием других культур ✓
5) Исследовать точность модели за пределами районов обучения

Заключение

- Произведено 2 попытки улучшить данные для обучения.
- Исследована территориальное распределение качества работы моделей
- Полученные модели превосходят существующий способ прогнозирования и имеют возможность давать прогноз в любое время сезона.
- Модели, анализирующие весь временной ряд превосходят существующий способ прогнозирования и имеют возможность давать прогноз в любое время сезона.
- Получена возможность создавать нейронные сети для множества разных культур, для которых имеется статистика в БД ПМО.