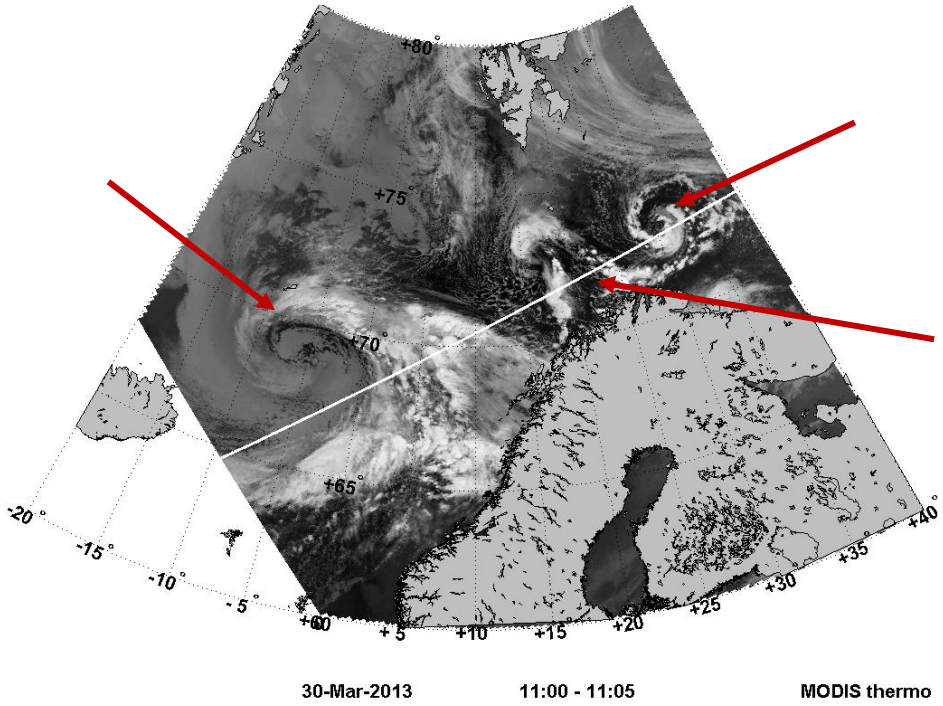
Применение методов машинного обучения для идентификации полярных мезомасштабных циклонов в данных численного моделирования атмосферы

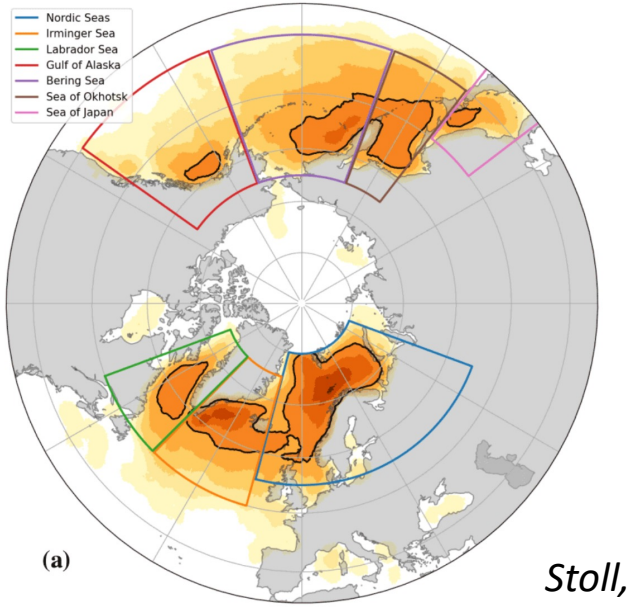
Левковская Юлия Алексеевна

Объект исследования



Циклонические вихри над поверхностью моря

Короткоживущие	6 – 36 часов
Быстро возникающие	3-6 часов время зарождения
Мезомасштабные	Диаметр - 200-1000 км
Интенсивные	Приводная скорость ветра >15 м/с



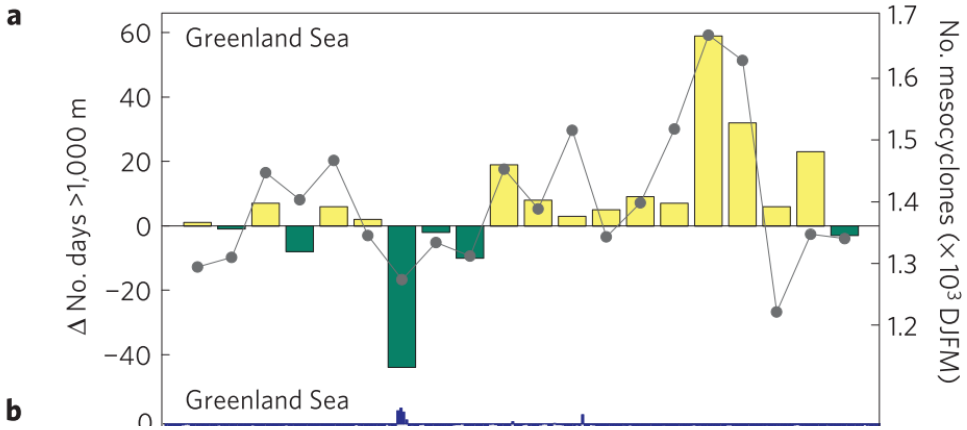
Stoll, 2022

- Распространены в Северном и южном полушарии выше 60° широты
- Возникают в зимние месяцы
- Более интенсивны в Северном полушарии

Актуальность

1. Малый размер и взрывной характер возникновения усложняют прогноз и анализ
2. Опасное метеорологическое явление для разных секторов экономики
3. Оказывают большое влияние на характеристики океана, а частности – глубокую конвекцию
4. Увеличение площади открытой воды в Арктике ведет к увеличению количества ПМЦ

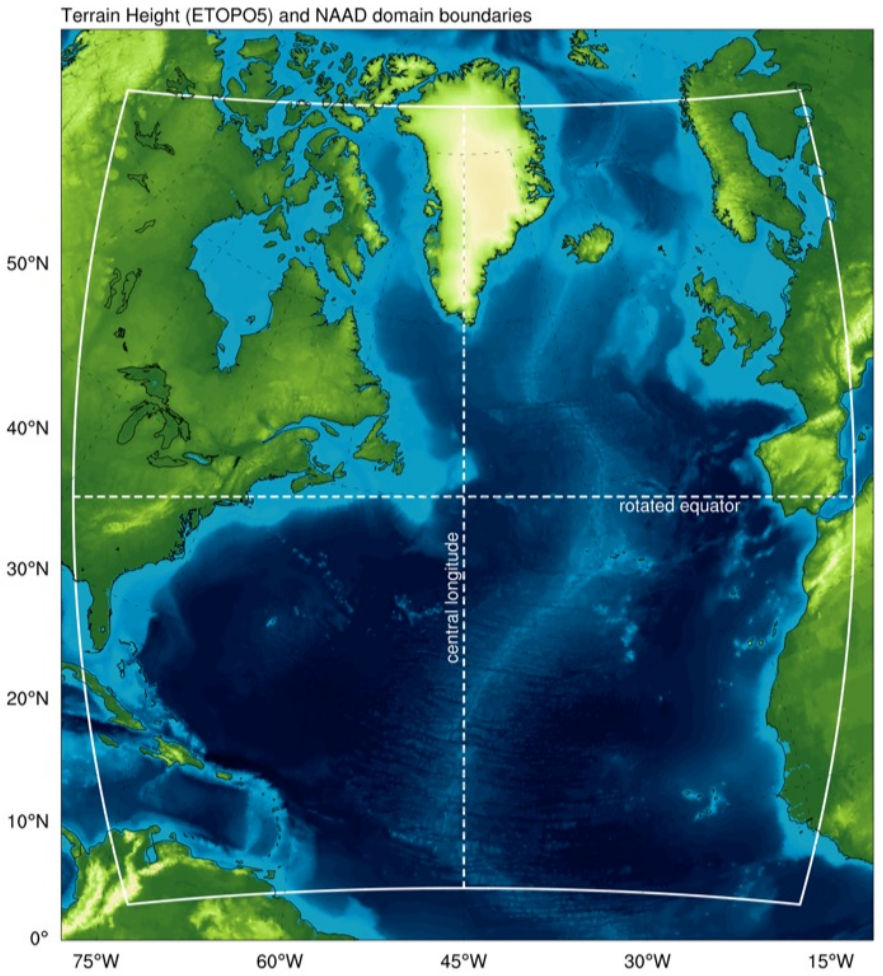
График количества полярных мезоциклонов и глубины конвекции в Гренландском море



Condron and Renfrew, 2013

Ретроспективный анализ RAS NAAD

КОНФИГУРАЦИЯ	LoRes	HiRes
Модель	WRF-ARW 3.8.1	
Приближение	гидростатика	негидростатика
Горизонтальное разрешение	**77 км**	**14 км**
Вертикальные уровни	50 (от 10 м до 50 гПа)	
RK3 шаг по времени	360 с	30 с
Частота записи	3 часа	3 часа

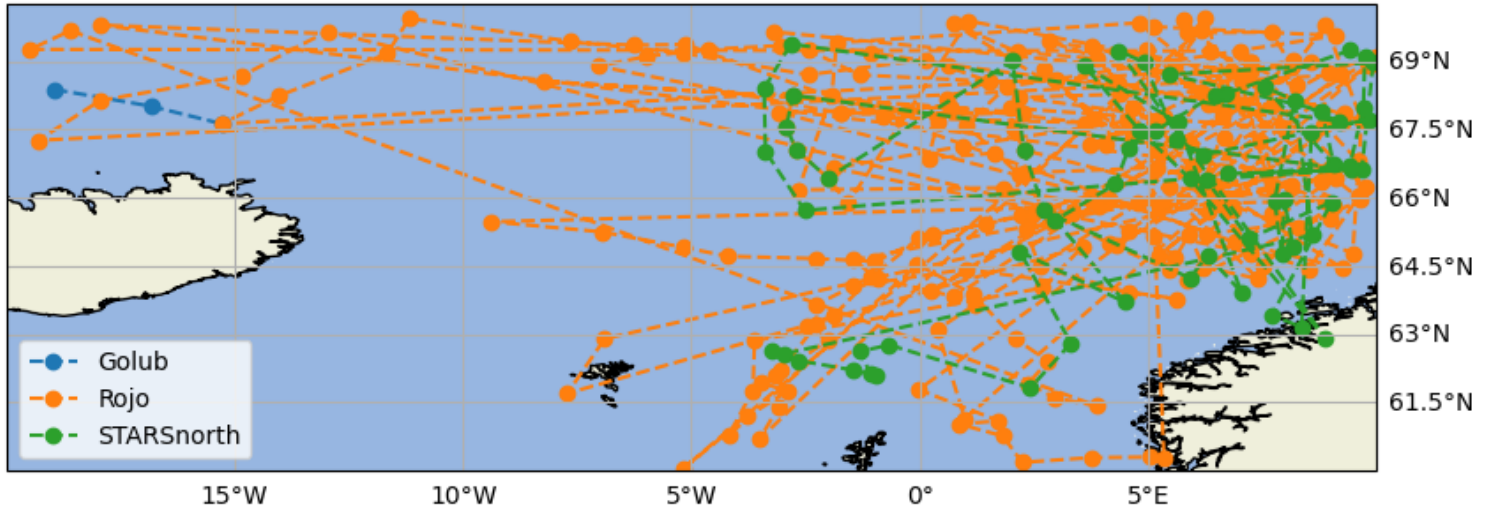


Gavrikov et. al., 2020

База данных ПМЦ

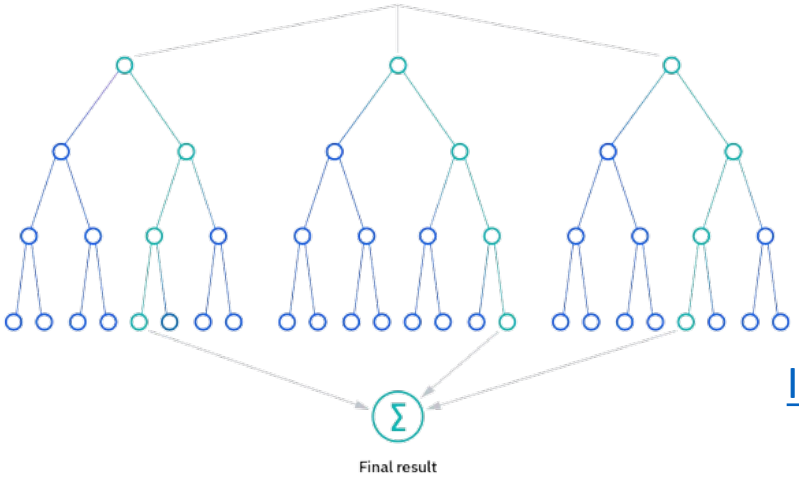
time	num	source	diam	cand_lat	cand_lon	selection	lat_sat	lon_sat
2000-03-08 18:00:00	4.0	Rojo	340.0	69.30325	17.364119	1.0	69.2908163265306	13.7091836734694
2000-03-08 18:00:00	4.0	Rojo	340.0	66.889114	10.077933	2.0	69.2908163265306	13.7091836734694
2000-03-08 21:00:00	4.0	Rojo	317.137546468401	68.643326	16.21061	1.0	68.314126394052	13.314126394052
2000-03-08 21:00:00	4.0	Rojo	317.137546468401	66.412155	10.47754	2.0	68.314126394052	13.314126394052
2000-03-09 00:00:00	4.0	Rojo	283.68029739777	68.37703	15.019739	1.0	67.3104089219331	12.3104089219331
2000-03-09 00:00:00	4.0	Rojo	283.68029739777	65.654144	10.260723	2.0	67.3104089219331	12.3104089219331
2000-03-09 03:00:00	4.0	Rojo	250.223048327138	64.3625	11.038791	2.0	66.3066914498141	11.3066914498141
2000-03-09 03:00:00	4.0	Rojo	250.223048327138	67.62386	14.18565	3.0	66.3066914498141	11.3066914498141
2001-04-09 06:00:00	12.0	Rojo	381.584158415842	73.82052	-8.084577	1.0	73.5	-9.0
2001-04-09 09:00:00	12.0	Rojo	407.180527383367	73.816795	-8.972105	1.0		
2001-04-09 12:00:00	12.0	Rojo	418.133874239351	73.86857	-9.307659	1.0		
2001-04-09 15:00:00	12.0	Rojo	429.087221095335	73.43687	-9.283808	1.0		

(Rojo et al., 2019), (Golubkin et al., 2021), (Noer et al., 2011)



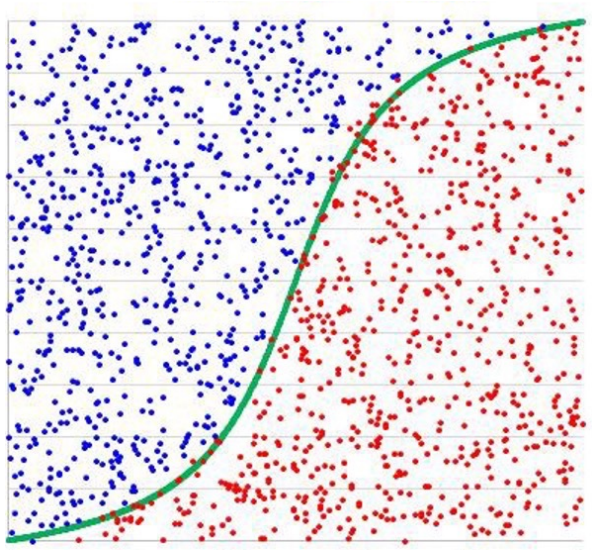
Модели машинного обучения

Случайный лес



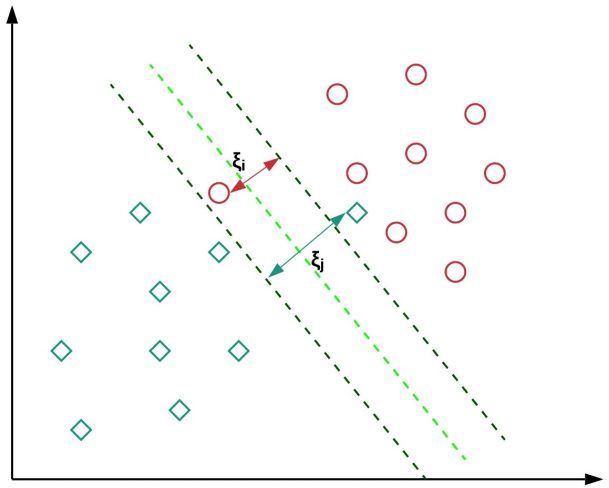
IBM

Логистическая регрессия



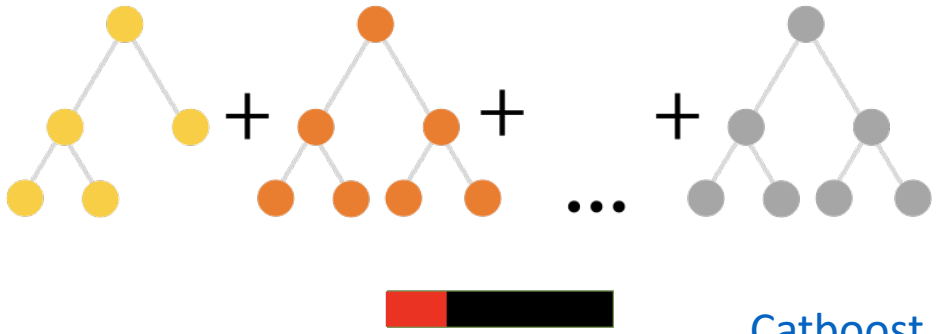
SF Education

Метод опорных векторов



Towards data science

CatBoost (градиентный бустинг)

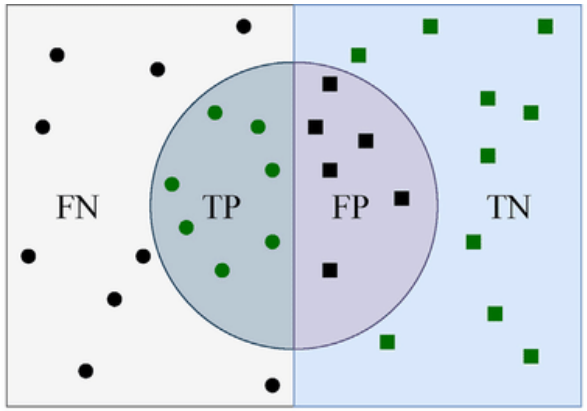


Catboost.ai

Метрики

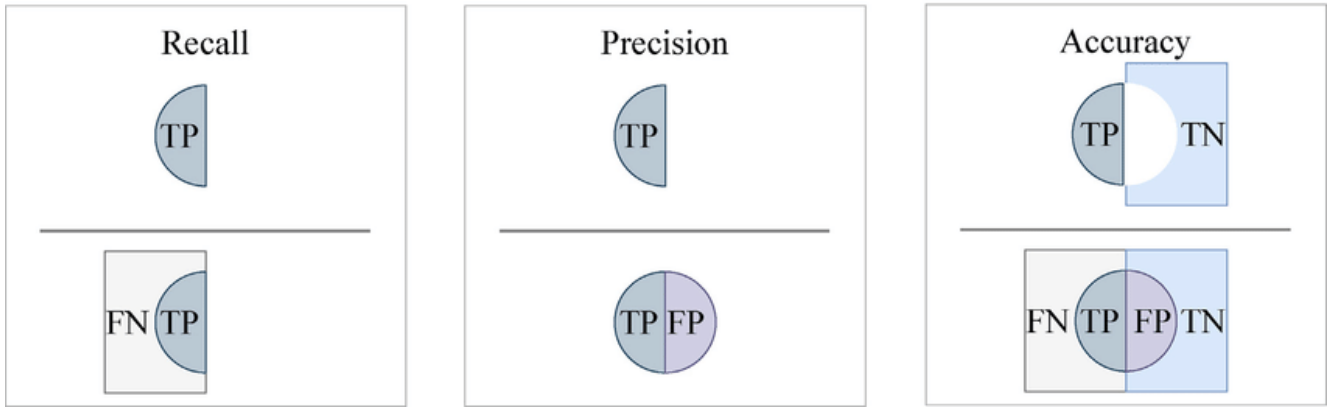
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$



$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

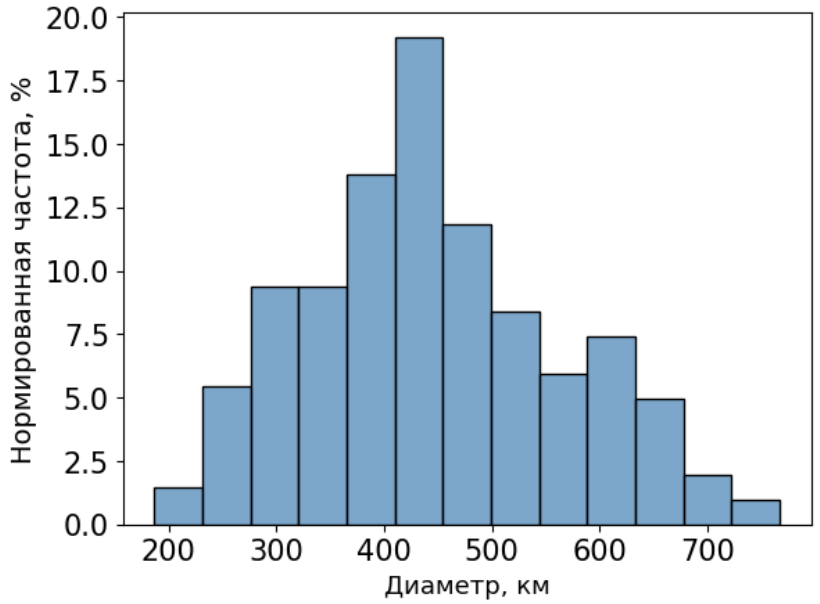
$$recall = \frac{TP}{TP + FN}$$

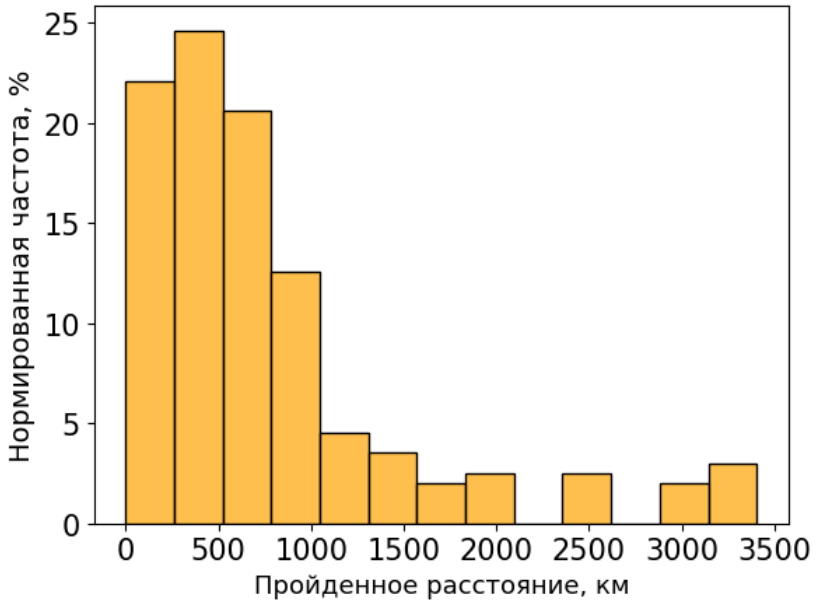


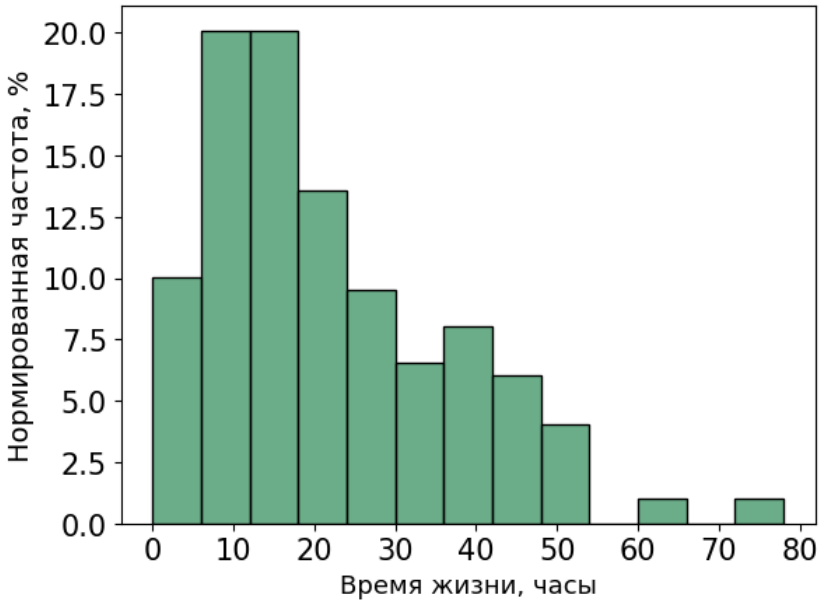
(Maleki,2020)

Распределения характеристик мезомасштабных циклонов

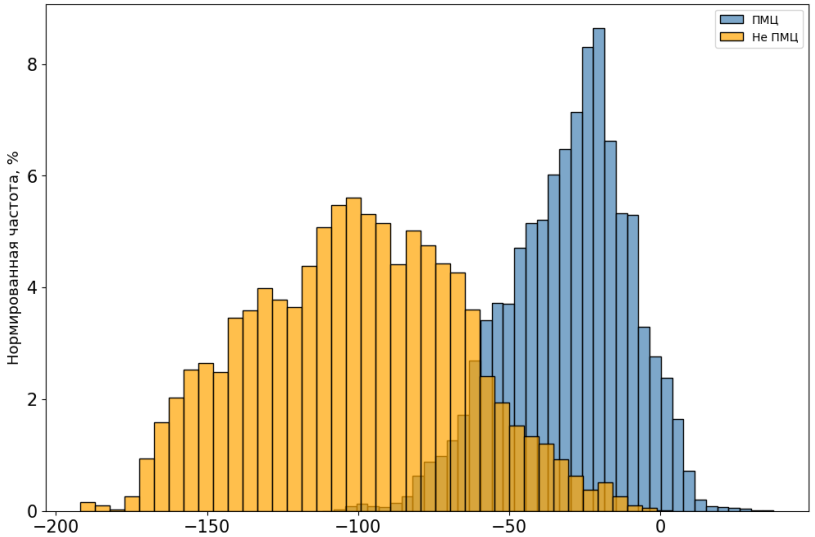
Составлена выборка из 116 мезомасштабных циклонов за период 2000-2015 и составлена выборка случайных областей, где мезомасштабные циклоны не наблюдались аналогичного размера



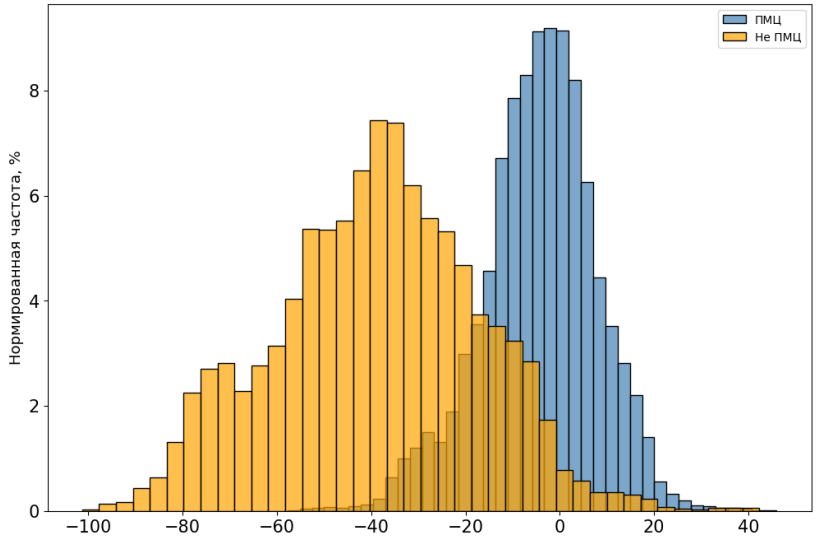




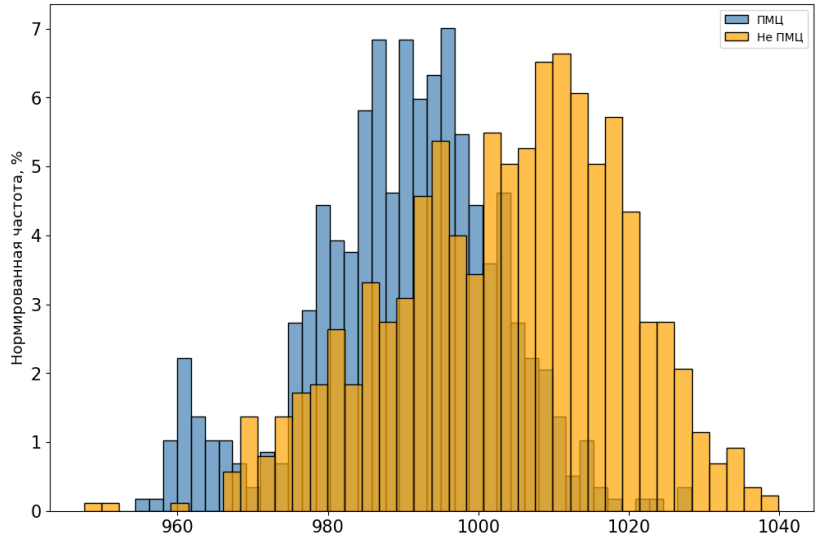
Распределения метеорологических характеристик



Индекс холодных вторжений на высоте 500 гПа



Индекс холодных вторжений на высоте 700 гПа



Минимальное давление на уровне моря

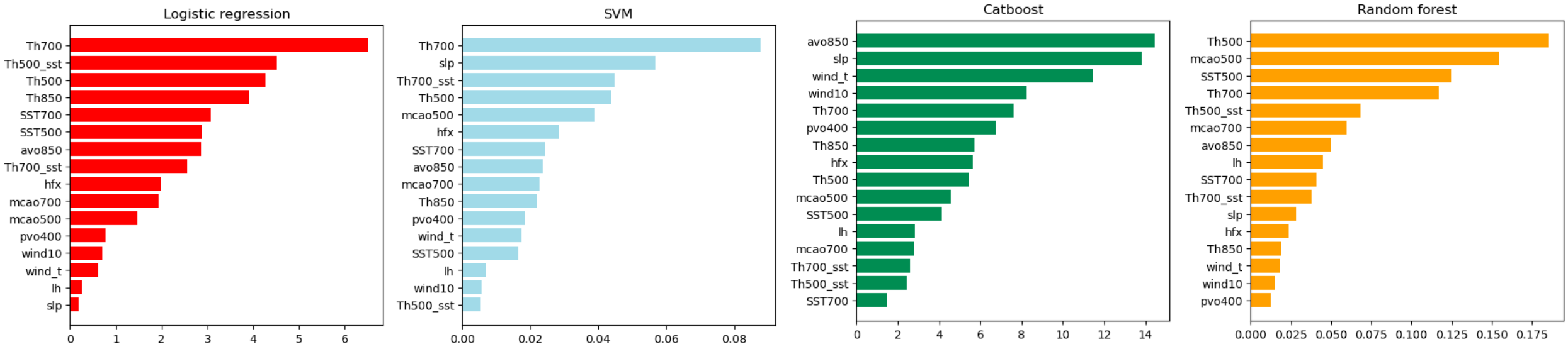
Результаты обучения моделей. Отложенная выборка

	Accuracy		Precision		Recall		F1	
	Трейн	Тест	Трейн	Тест	Трейн	Тест	Трейн	Тест
Логистическая регрессия	0.964	0.977	0.967	0.983	0.958	0.972	0.975	0.977
Модель опорных векторов	0.998	0.971	0.995	0.972	1	0.972	0.975	0.972
Случайный лес	1	0.977	1	0.972	1	0.983	1	0.978
CatBoost	1	0.974	1	0.967	1	0.983	1	0.975

Результаты обучения моделей. Кросс-валидация

	Accuracy		Precision		Recall		F1	
	Трейн	Тест	Трейн	Тест	Трейн	Тест	Трейн	Тест
Логистическая регрессия	0.967	0.959 ± 0.06	0.969	0.96 ± 0.1	0.964	0.959 ± 0.08	0.966	0.959 ± 0.05
Модель опорных векторов	0.998	0.971 ± 0.04	0.996	0.961 ± 0.07	1	0.983 ± 0.05	0.998	0.971 ± 0.03
Случайный лес	1	0.979 ± 0.05	1	0.971 ± 0.09	1	0.988 ± 0.03	1	0.979 ± 0.04
CatBoost	1	0.983 ± 0.04	1	0.975 ± 0.09	1	0.991 ± 0.03	1	0.982 ± 0.05

Значимость признаков на кросс-валидации

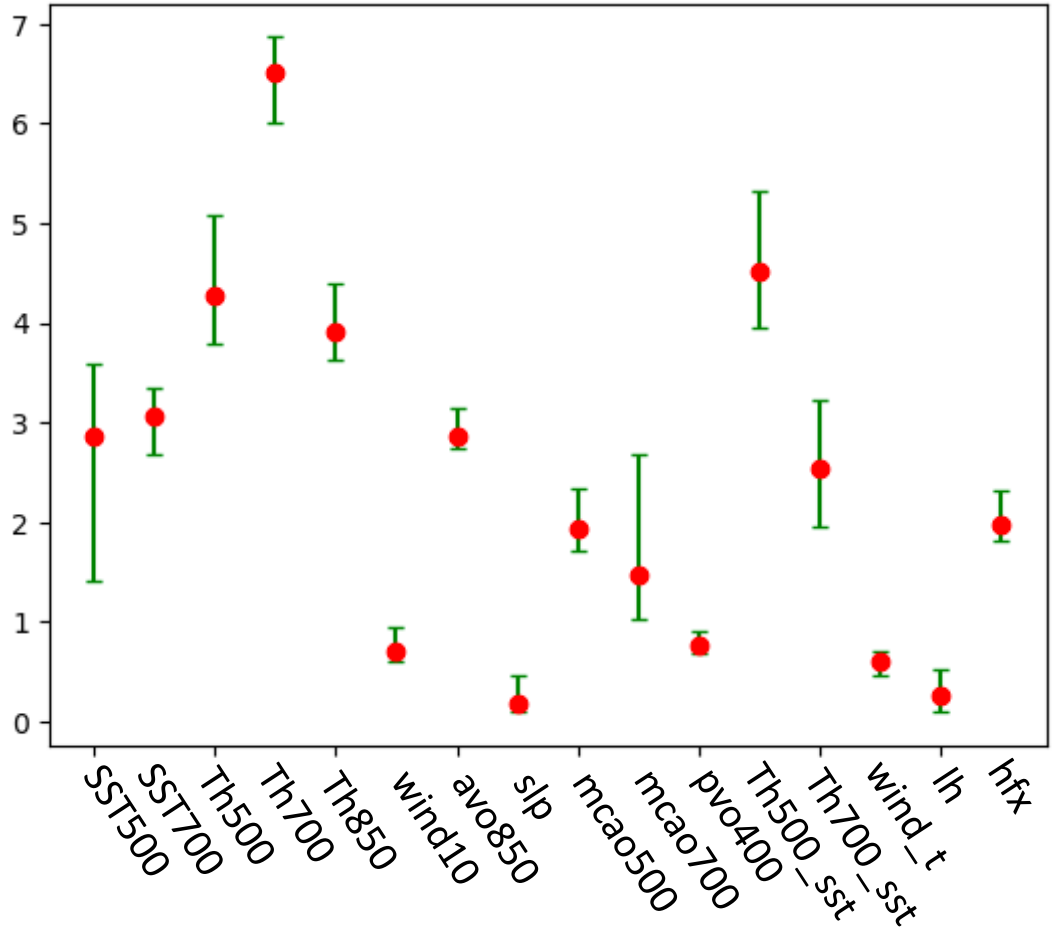


Наиболее значимые признаки в логистической регрессии – Th700, Th500_sst, Th500, в модели опорных векторов – Th700, slp, Th700_sst, в модели CatBoost – avo850, slp, wind_t, в модели случайного леса – Th500, mcao500, SST500

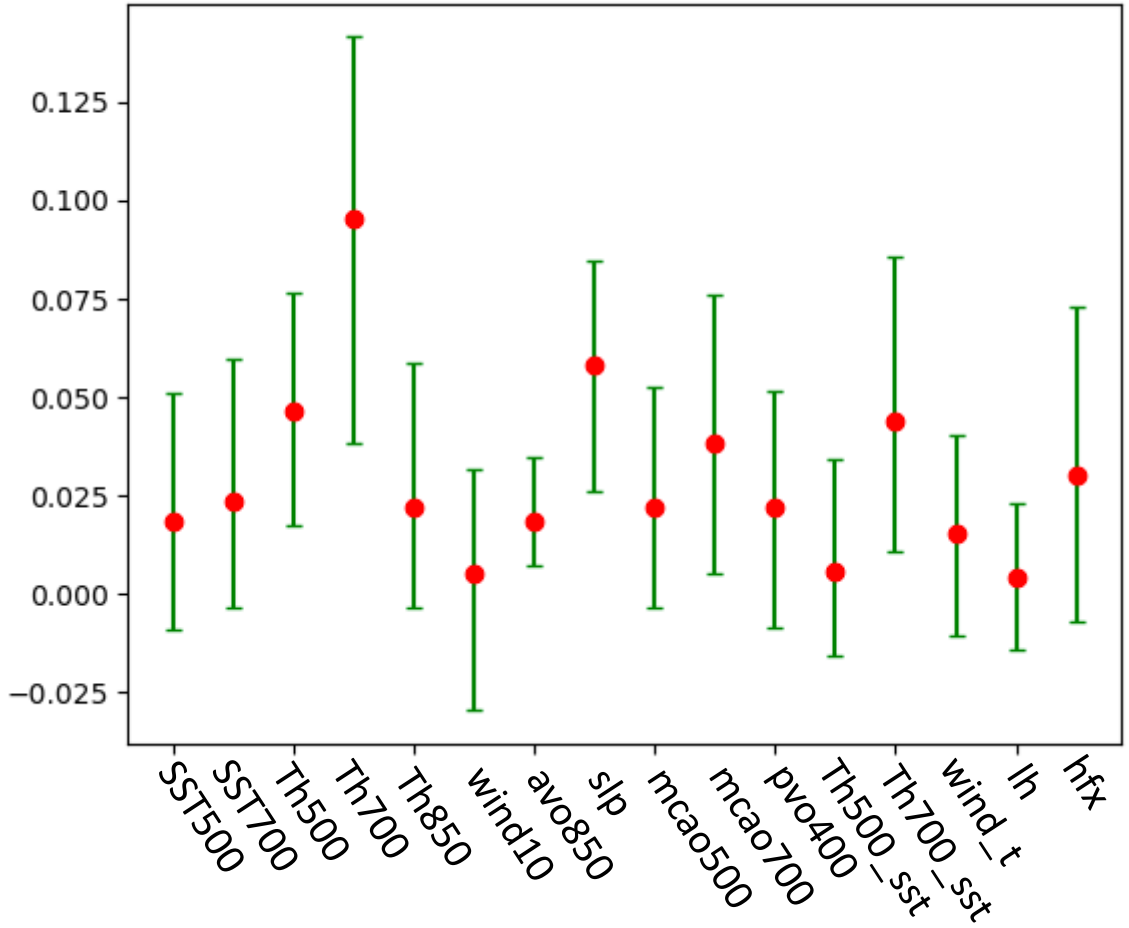
Таким образом наблюдается сильная несогласованность в значимости признаков между моделями

Разброс значимости признаков

Логистическая регрессия

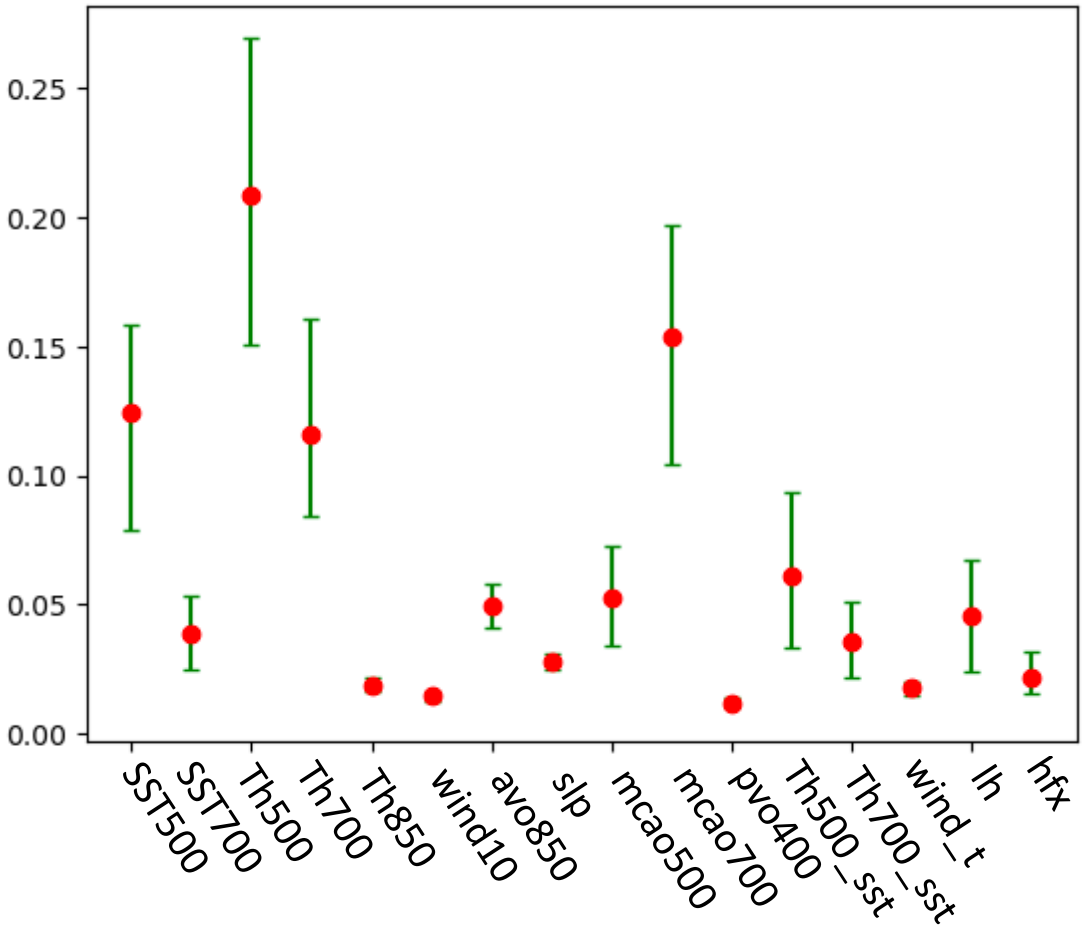


Модель опорных векторов

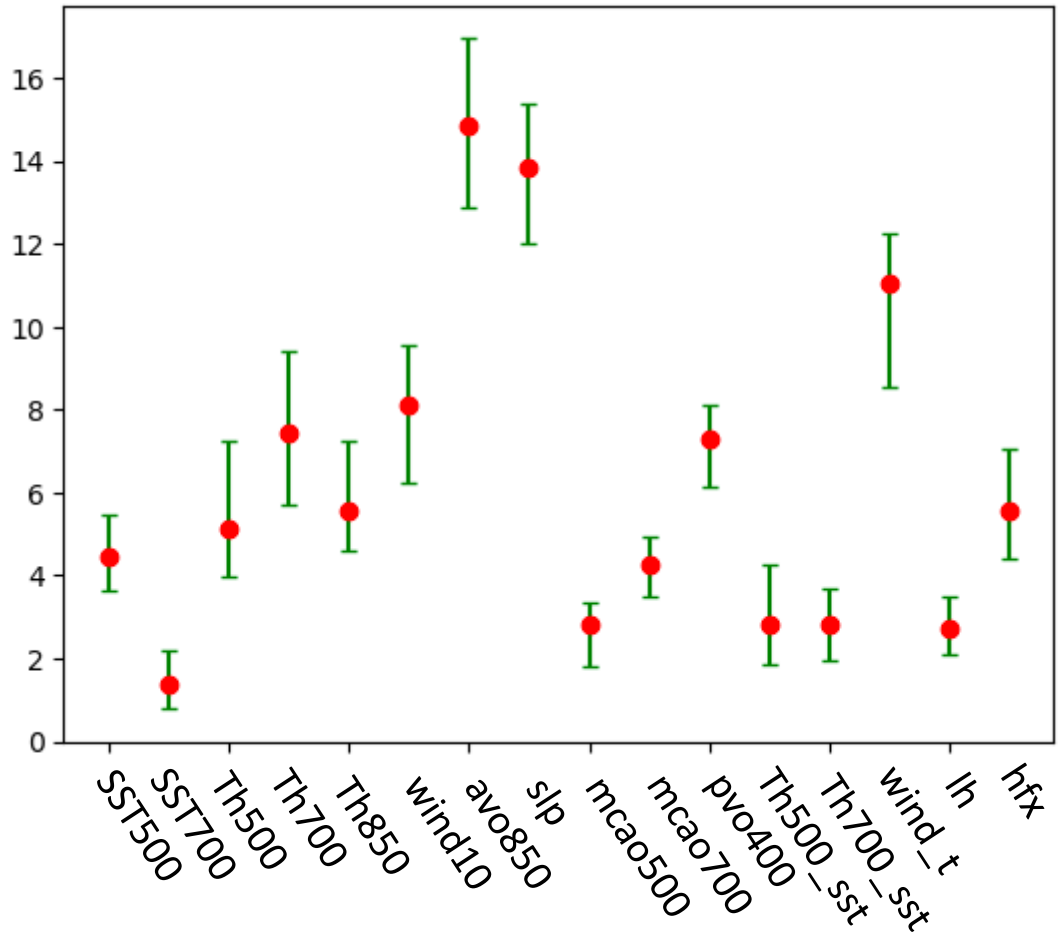


Разброс значимости признаков

Случайный лес



CatBoost

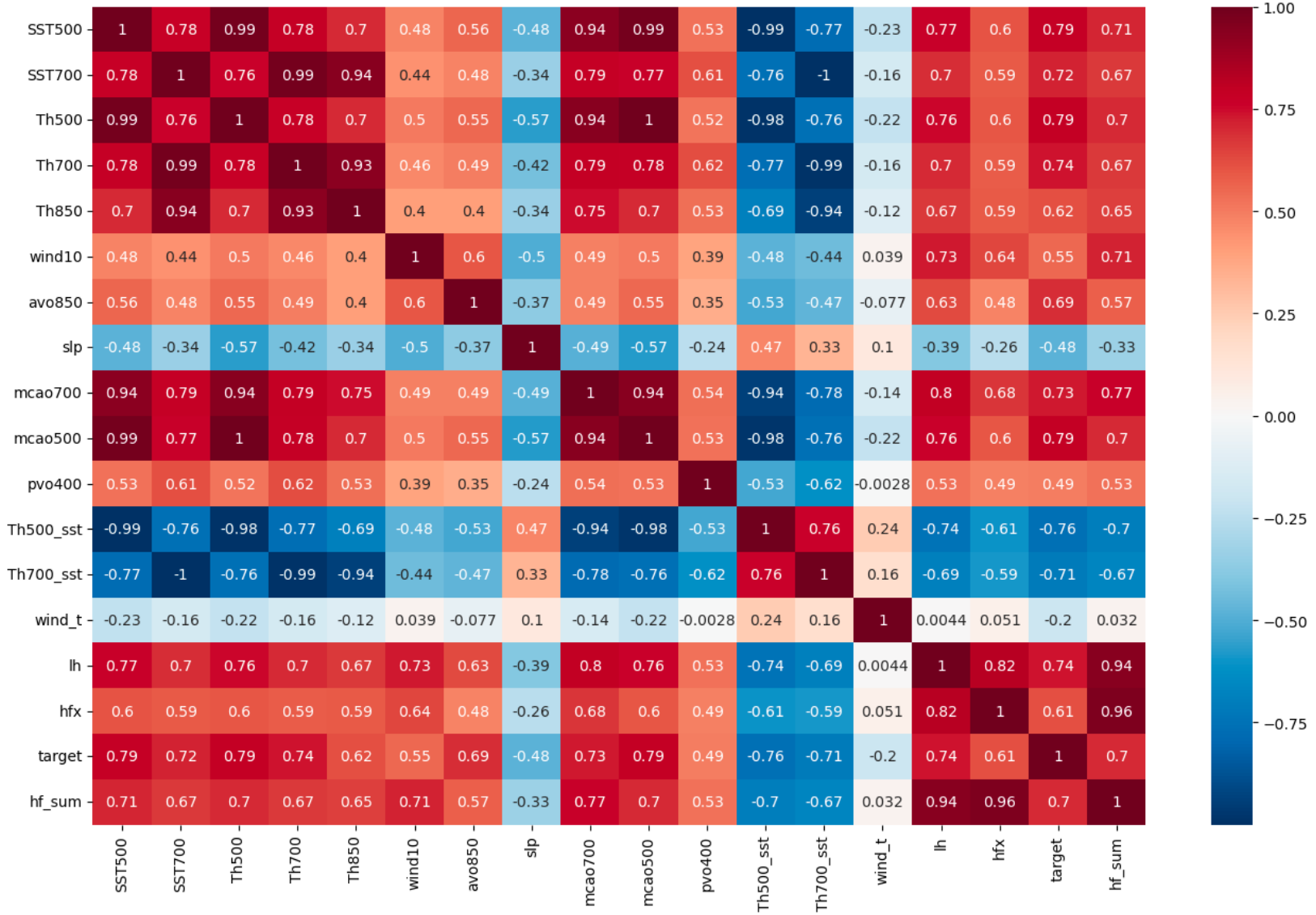


Корреляция признаков

Между многими признаками наблюдается высокая или очень высокая корреляция.

Вероятнее всего это стало причиной сильной несогласованности наиболее значимых признаков в разных моделях

Было решено отобрать наименее скоррелированные признаки и поэтапно добавлять к ним признаки из числа сильно скоррелированных

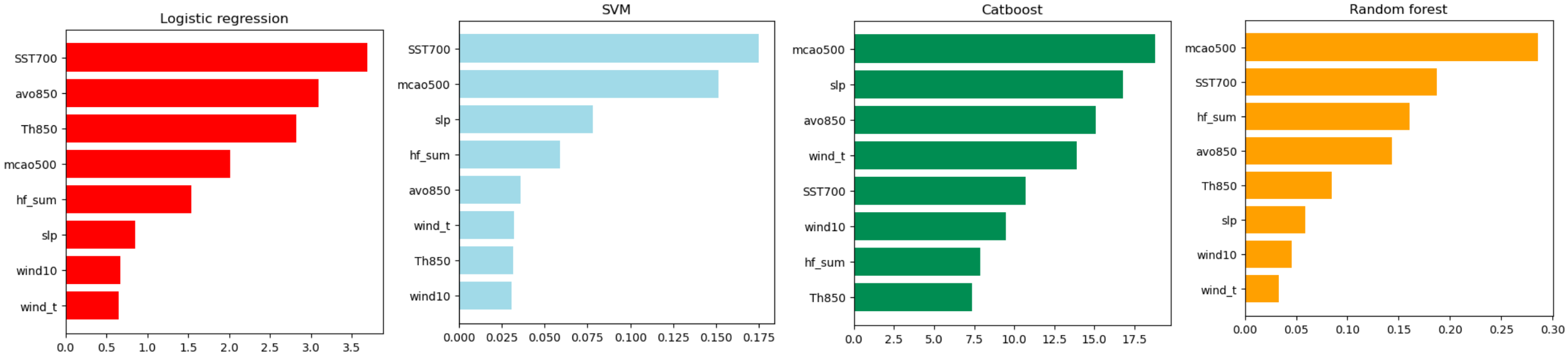


Результаты обучения моделей на новых признаках

	Accuracy		Precision		Recall		F1	
	Трейн	Тест	Трейн	Тест	Трейн	Тест	Трейн	Тест
Логистическая регрессия	0.961	0.96 ± 0.08	0.963	0.962 ± 0.07	0.959	0.959 ± 0.08	0.961	0.959 ± 0.06
Модель опорных векторов	0.998	0.981 ± 0.03	0.995	0.973 ± 0.07	1	0.99 ± 0.05	0.998	0.981 ± 0.03
Случайный лес	1	0.983 ± 0.04	1	0.98 ± 0.07	1	0.988 ± 0.03	1	0.983 ± 0.04
CatBoost	1	0.983 ± 0.04	1	0.979 ± 0.07	1	0.986 ± 0.05	1	0.982 ± 0.04

В таблице представлено качество модели, обученной на наименее скоррелированных признаках (slp, avo850, pvo400,wind10, wind_t, hfx_sum) и трех признаках из числа наиболее скоррелированных (mcao500, SST700, Th850)

Значимость новых признаков в моделях



Наиболее значимые признаки в логистической регрессии – SST700, avo850, Th850, в модели опорных векторов – SST700, mcao500, slp, в модели Catboost – mcao500, slp, avo850, в модели случайного леса – mcao500, SST700, hf_sum

mcao500 и SST700 вошли в три самых значимых признака в трех из четырех моделей, slp и avo850 - в двух из четырех

Выводы

- Наилучший результат показали модели машинного обучения, обученные на некоррелированных признаках с добавлением трех из числа коррелированных
- Самыми значимыми признаками являются: индекс холодных вторжений на 500 гПа, разница между температурой на высоте 700 гПа и на уровне моря, завихренность на 850 гПа и давление на уровне моря
- Состав выбранных признаков хорошо согласуется с физическими свойствами мезомасштабных циклонов и характеризует их основные отличительные особенности: пониженное давление на уровне моря и сильную вертикальную конвекцию
- Модели МО показали хороший результат в задаче идентификации мезомасштабных циклонов по всем выбранным метрикам


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •