Оценка бонитета и возраста лесных насаждений с помощью данных ДЗЗ и методов машинного обучения

Хвостиков С.А., Барталев С.А.
khvostikov@d902.iki.rssi.ru

Институт Космических Исследований РАН

Работа выполнена в рамках реализации важнейшего инновационного проекта государственного значения

"Разработка системы наземного и дистанционного мониторинга пулов углерода и потоков парниковых газов на территории Российской Федерации, обеспечение создания системы учета данных о потоках климатически активных веществ и бюджете углерода в лесах и других наземных экологических системах» (рег. № 123030300031-6).

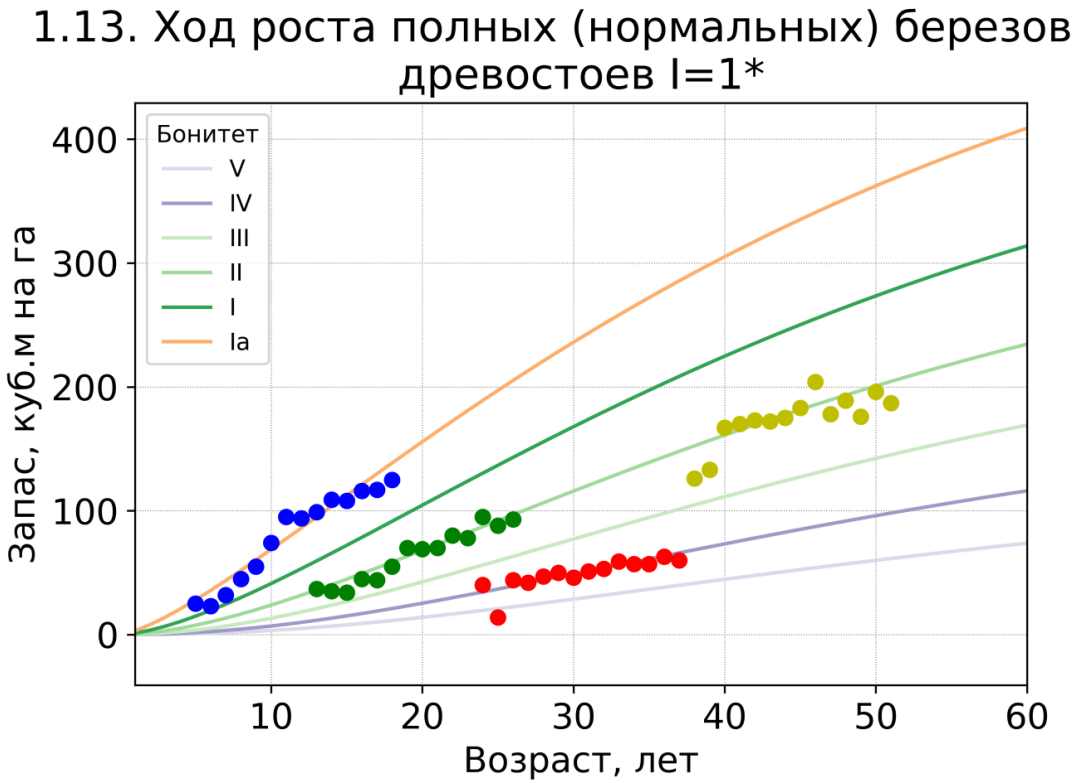
Двадцать первая международная конференция
«Современные проблемы дистанционного зондирования Земли из космоса»
ИКИ РАН 13 – 17 ноября 2023 г.

Введение

Данные о продуктивности и возрасте лесных насаждений важны для решения задач лесного хозяйства, а также для оценки запаса и потока углерода лесов;

В данной работе проведены исследования по выбору признаков для оценки этих характеристик;

Также проведено сравнение эффективности разных методов машинного обучения при решении этой задачи.



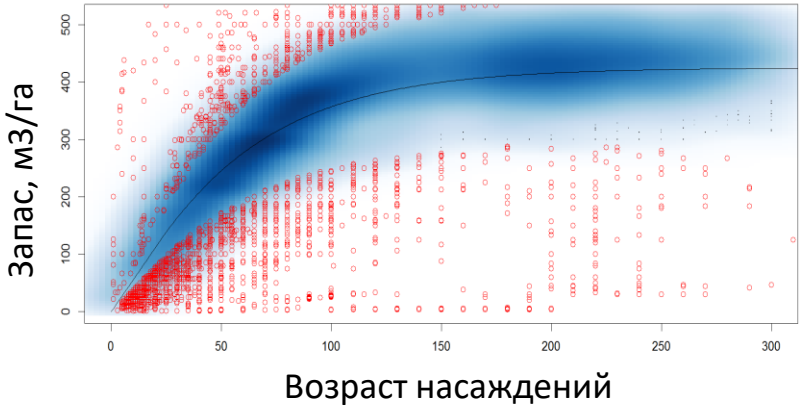
Обучающая выборка

Обучающая выборка основана на данных таксационных выделов;

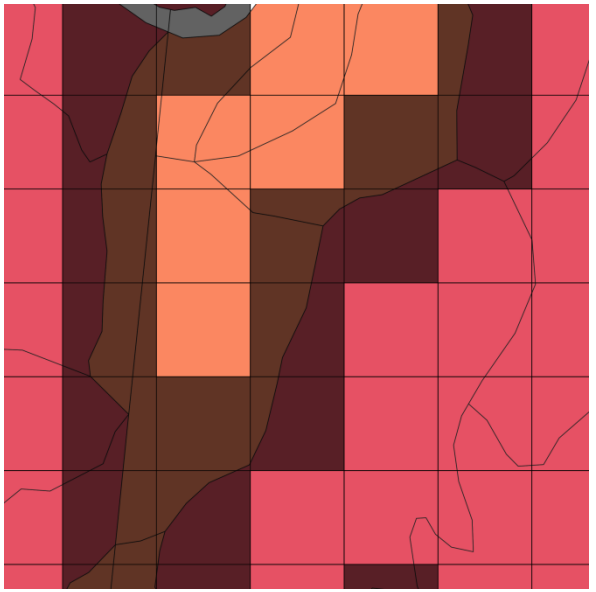
Данные огрубляются на уровень пикселов MODIS (230 м), с фильтрацией неоднородных пикселов и пикселов со сбойными наблюдениями;

Всего было собрано более 28 миллионов пикселов с информацией о бонитете и возрасте, использованных для обучения и валидации.

Фильтрация выборки на основе моделей хода роста и данных о возрасте, бонитете, породе и запасе насаждений



Выбор однородных выделов на уровне пикселов MODIS



Признаки

Предполагается, что бонитет и возраст определяют продуктивность леса, и для их оценки необходимы данные многолетних наблюдений. Почти все используемые данные даны на период с 2001 по 2022 годы.

Используются интерполированные данные MODIS за конец весны, лето, начало осени (Миклашевич и др, 2021).
Используются продукты, полученные в ИКИ РАН – запас (Ворушилов и др., 2020), полнота и лесистость (Ховратович 2022), преобладающая порода (Жарко и др 2023).
Для всех признаков были построены средние, медианы, вариация и тренд за 2001-2022 годы.
Также используется карта рельефа ASTER.
Всего было собрано 665 признаков.

Жарко В.О., Барталев С.А., Богодухов М.А., Егоров В.А., Хвостиков С.А. Развитие методов спутникового картографирования преобладающих древесных пород в лесах всей территории России по данным MODIS // Материалы 21-й международная конференция "Современные проблемы дистанционного зондирования Земли из космоса". 2023.
Ховратович Т.С. Показатели горизонтальной структуры лесов и их дистанционная оценка на основе оптических спутниковых данных (лекция) // Материалы 20-й Международной конференции «Современные проблемы дистанционного зондирования Земли из космоса». 2022
Миклашевич Т.С., Барталев С.А., Егоров В.А. Развитие предварительной обработки данных спутниковых наблюдений приборов VIIRS и MODIS для задачи мониторинга растительного покрова // Материалы 19-й Международной конференции «Современные проблемы дистанционного зондирования Земли из космоса». 2021.
Ворушилов И.И., Барталев С.А., Егоров В.А. Развитие метода оценки запасов стволовой древесины с использованием данных Terra-MODIS // Материалы Восемнадцатой Всероссийской Открытой конференции «Современные проблемы дистанционного зондирования Земли из космоса. 2020

Подход исследования

Большой объем выборки и большой набор признаков затрудняет построение регрессии и оценку бонитета и возраста;

С помощью различных методов оценки значимости переменных был выбран уменьшенный набор признаков;

Было проведено сравнение разных методов машинного обучения на выбранном наборе признаков;

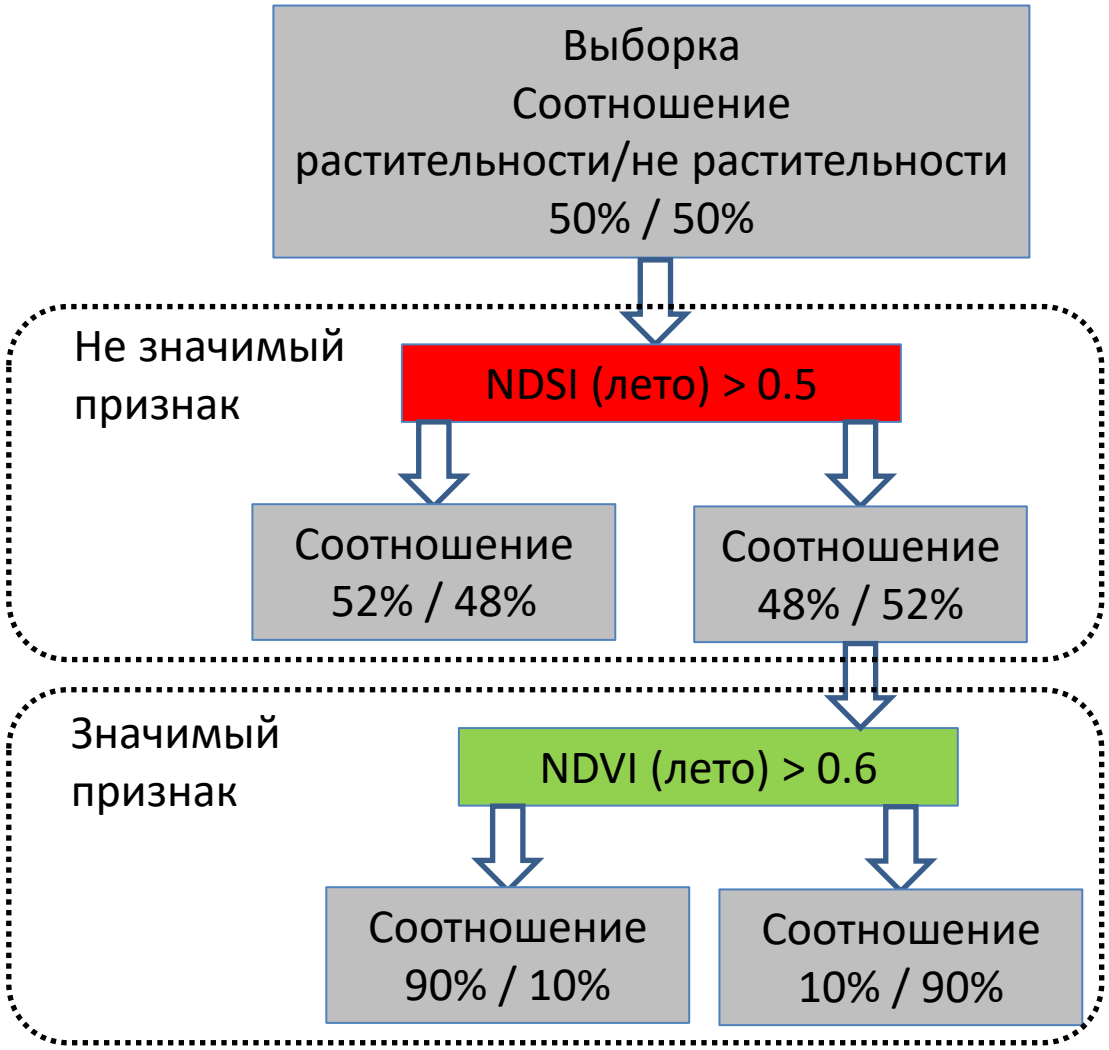
В конце новый набор признаков и оптимальный метод были применены на большой выборке для получения итоговых карт бонитета и возраста.

Методы оценки значимости, деревья решений

Случайные леса и другие популярные методы на основе деревьев решений используют только часть признаков в каждом узле;

Значимость признака можно определить по тому, насколько хорошо признак способен разбить исходную выборку;

Агрегация такой статистики по большому количеству узлов и деревьев может дать адекватную итоговую оценку значимости.

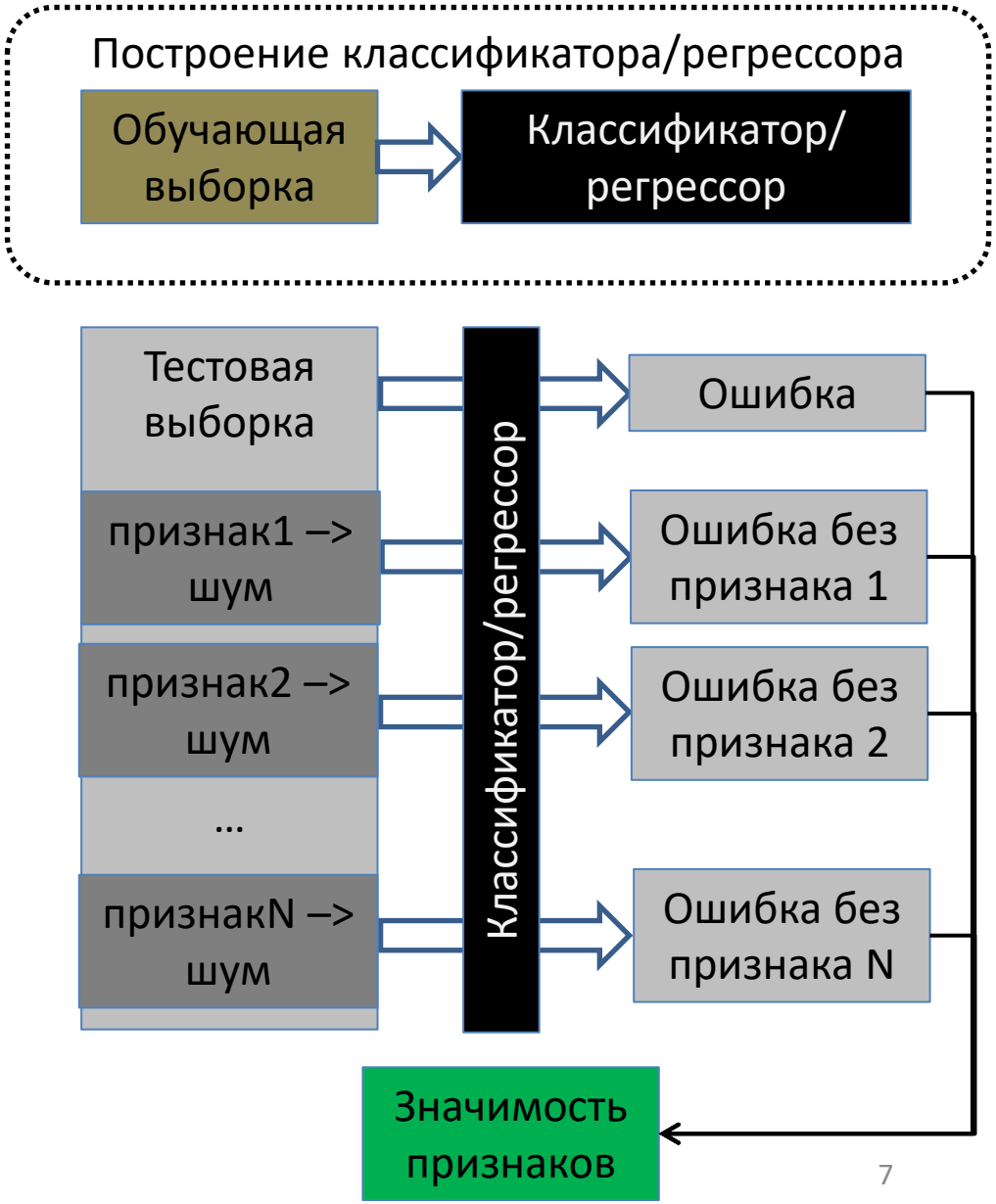


Методы оценки значимости, изменение признаков

Другой подход оценки значимости основан на использовании уже готового классификатора;

На его вход можно подать тестовую выборку, в которой один признак заменен на шум, и оценить падение точности;

Агрегация падений точности при подмене признака позволяет оценить значимость признаков.



Использованные методы оценки значимости

Для оценки значимости используются 6 методов из двух библиотек – ranger (случайные леса) и LGBM (градиентный бустинг).

Информативность признака в узлах дерева
ranger_impurity
LGBM_split
LGBM_gain

Замена признака на шум
ranger_permutation
ranger_outer
LGBM_outer

В качестве эталона использовалось итеративное построение регрессии по каждому признаку, выбор наилучшей модели и переход на следующую итерацию с добавлением следующего признака (model).

Оценка значимости признаков

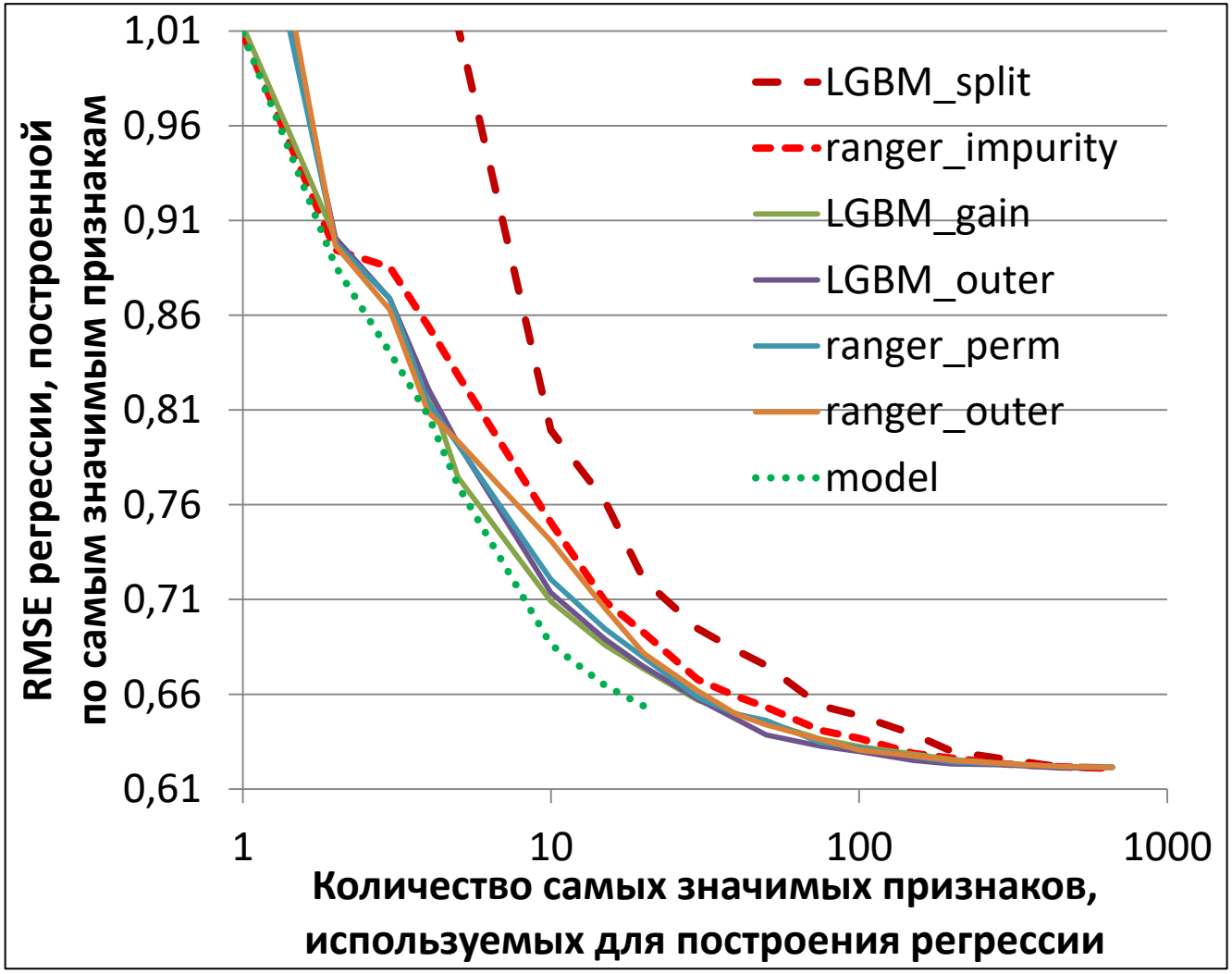
Для сравнение методов по самым значимым признакам строилась модель, ее точность характеризовала качество набора признаков.

LGBM_split – хуже всего
ranger_impurity – также плохо.

Модельный – лучше всего, но медленный.

Остальные – одинаковы. Выбран LGBM_outer

100 признаков
RMSE 0,629
665 признаков
RMSE 0,621



Подходы построения регрессии

Линейная регрессия – известный уже в течение 100 лет базовый метод;

Альтернативные методы построения линейной регрессии – Lasso, Ridge, Elastic Net. Применяют подходы регуляризации для более точного определения коэффициентов линейной регрессии;

Случайные леса (sklearn, skranger). Ансамбль деревьев решений, в котором на вход каждого дерева подается случайная подвыборка, а в каждом узле дерева она оптимальным образом разделяется по случайному набору признаков.

Подходы построения регрессии – градиентный бустинг

Градиентный бустинг применительно к деревьям предполагает, что каждое следующее дерево строится не независимо, а старается минимизировать ошибки предыдущих деревьев.

Было опробовано 2 реализации градиентного бустинга:

XGBoost – одна из самых популярных библиотек для градиентного бустинга.

LightGBM (LGBM) – отличается от XGBoost тем, что преобразует признаки, превращая их по сути в гистограмму. Также метод формирует деревья в глубину, а не в ширину.

Сравнение методов классификации

На выборке из 2 млн элементов и 100 лучших признаков были обучены разные методы регрессии;

Их точность оценивалась по отдельной выборке из 1 млн элементов;

Лучшим методом оказался LightGBM;

Метод опорных векторов и k ближайших соседей оказались слишком медленными на большой выборке с большим количеством признаков.

Метод		R2	RMSE	Время обучения, минут
Линейная регрессия	Базовая	0,65	0,81	0,05
	Ridge	0,65	0,81	0,05
	Lasso	0,65	0,81	2
	Elastic Net	0,65	0,81	2
К ближайших соседей		-	-	∞
Метод опорных векторов		-	-	∞
Случайный лес		0,776	0,65	75
XGBoost		0,783	0,64	5
LightGBM		**0,816**	**0,59**	**8**

Оценка значимости признаков, возраст

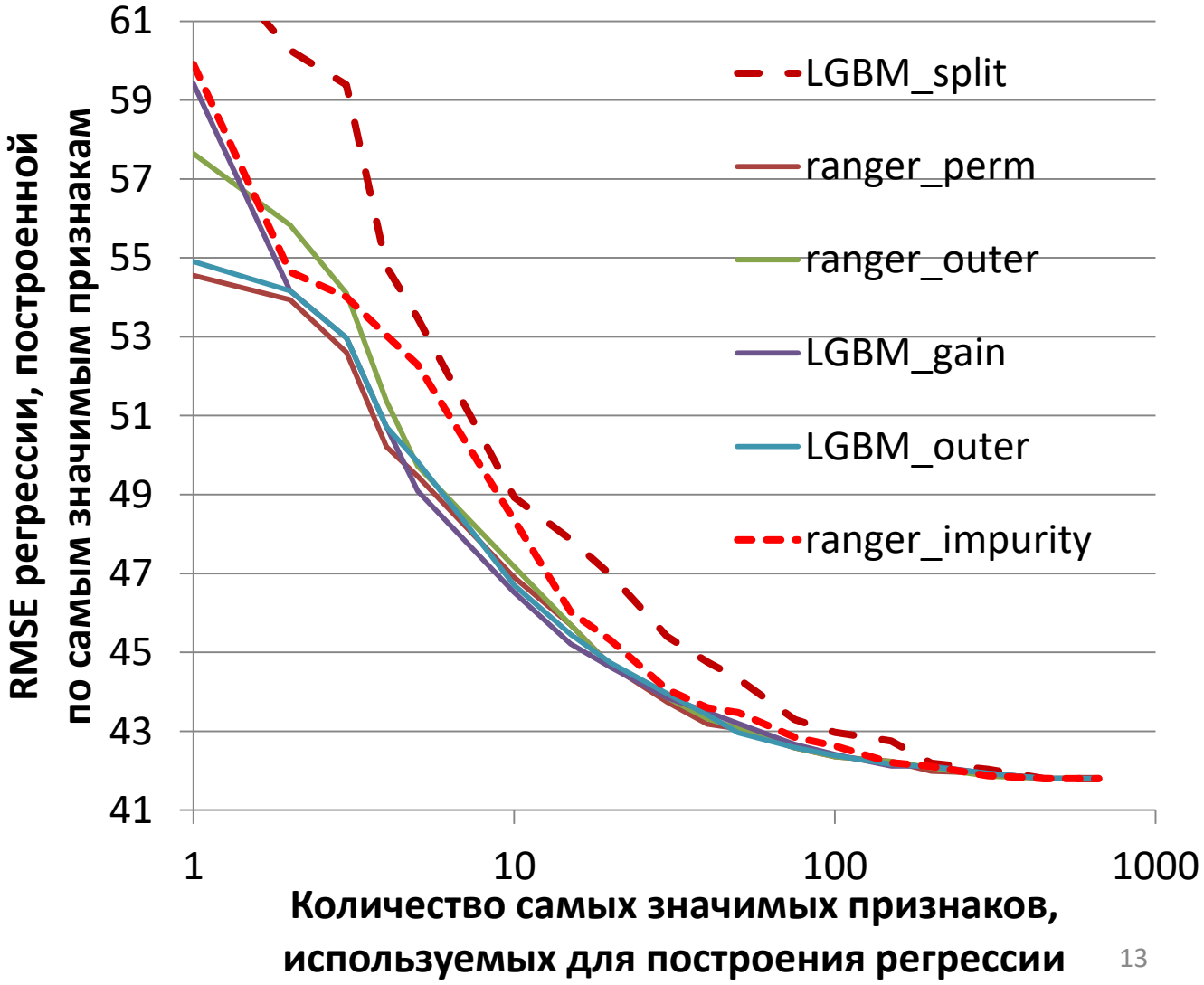
Для возраста разные методы оценки значимости дают тот же результат, что и для бонитета;

LGBM_split , ranger_impurity – также плохо.

Остальные – посередине. Выбран LGBM_outer

100 признаков
RMSE 42,4 года

665 признаков
RMSE 41,8 лет



Сравнение методов классификации, возраст

Как и ранее была взята обучающая и тестовая выборка из 2 млн и 1 млн элементов;

Лучшим методом опять оказался LGBM.

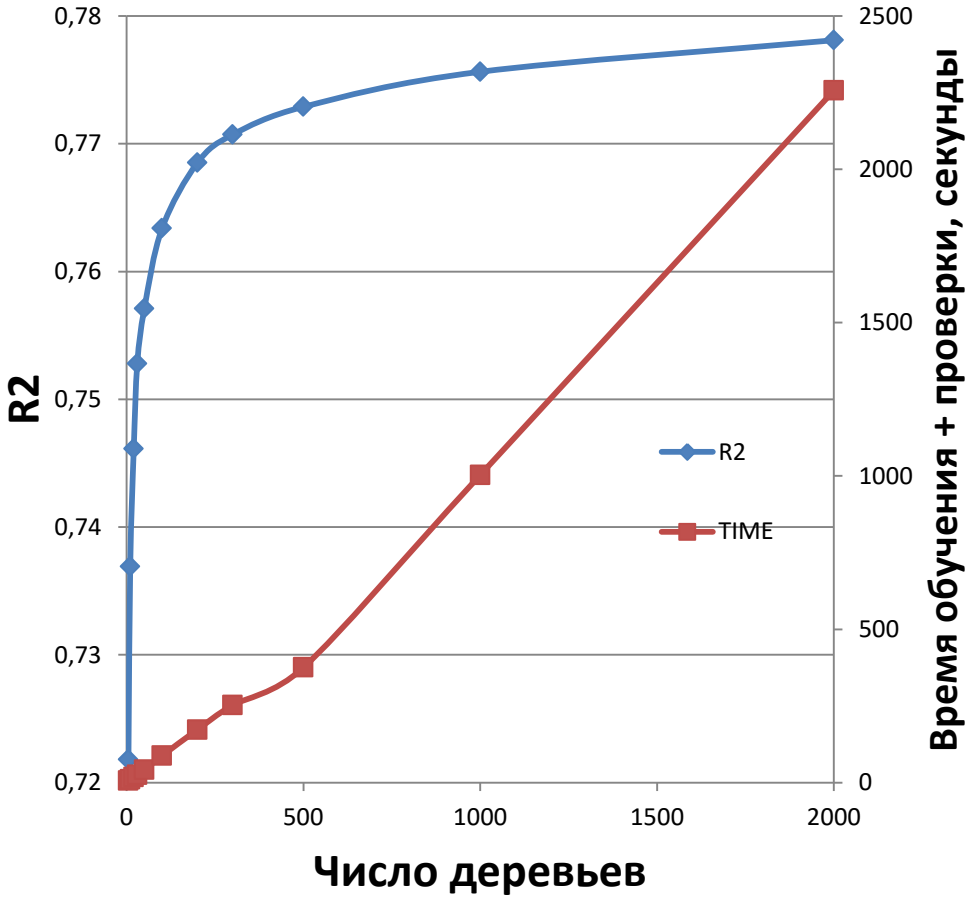
Метод	R2	RMSE	Время обучения, минуты
Линейная регрессия Elastic Net	0,4	48,8	2
Случайный лес	0,58	41	75
XGBoost	0,59	40,2	2,5
LightGBM	**0,64**	**37,8**	**7**

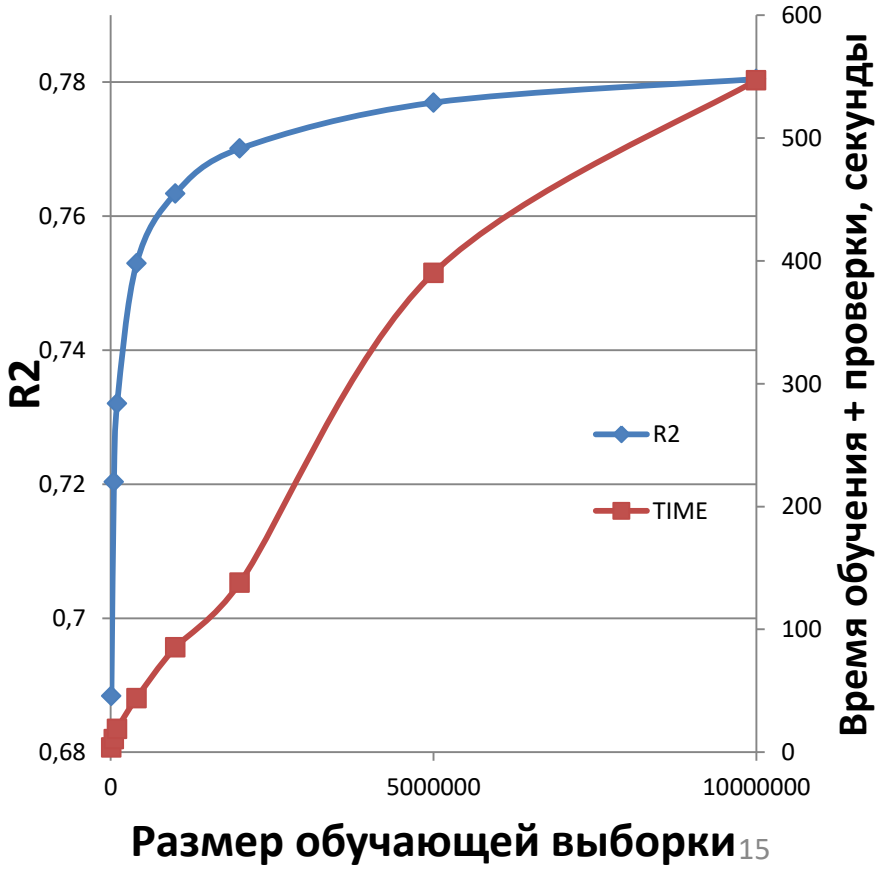
Анализ показывает, что погрешность зависит от породы лесных насаждений.

Порода	**RMSE**		**Порода**	**RMSE**
Кедр	41,9		Береза	26,5
Сосна	39,8		Липа	24,3
Бук	39,8		Осина	23,0
Пихта	35,9			

Настройка LGBM

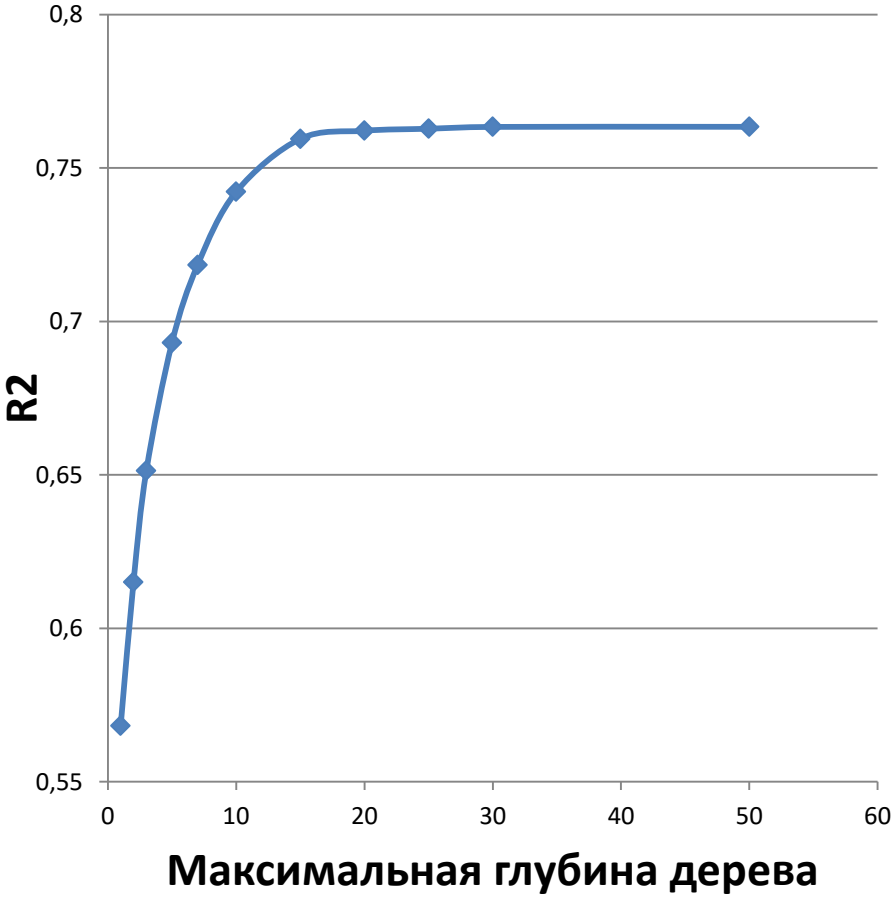
Точность LGBM сильно зависит от объема выборки, а также числа деревьев. Больше – лучше, но наблюдается насыщение.

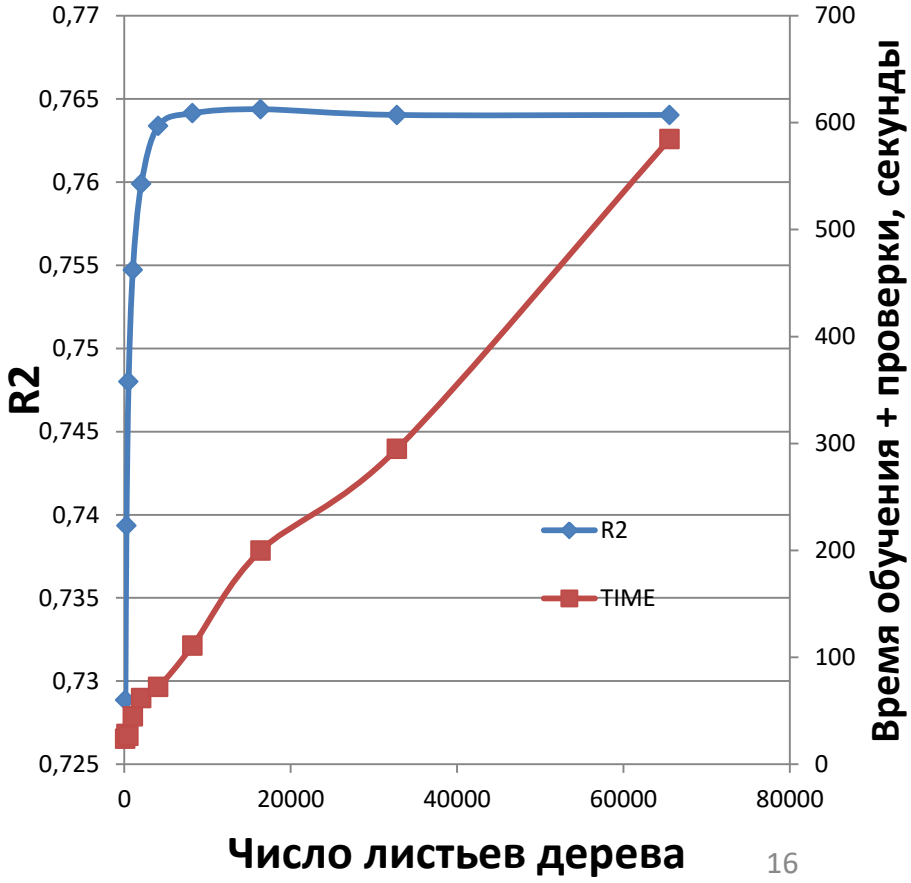




Настройка LGBM

Нужно установить число листьев (узлов) и глубину дерева, который в этом случае имеют некоторый оптимальный порог.
В теории большие значения могут весть к переобучению, но этот эффект слабо проявился при оценке бонитета.

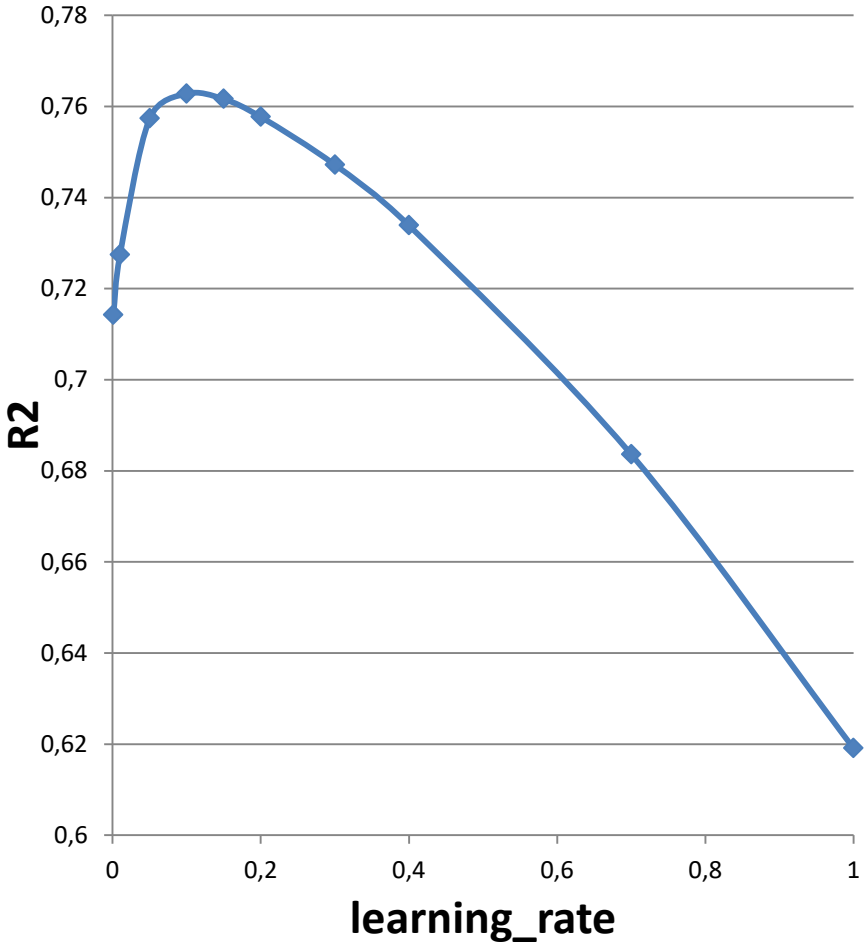


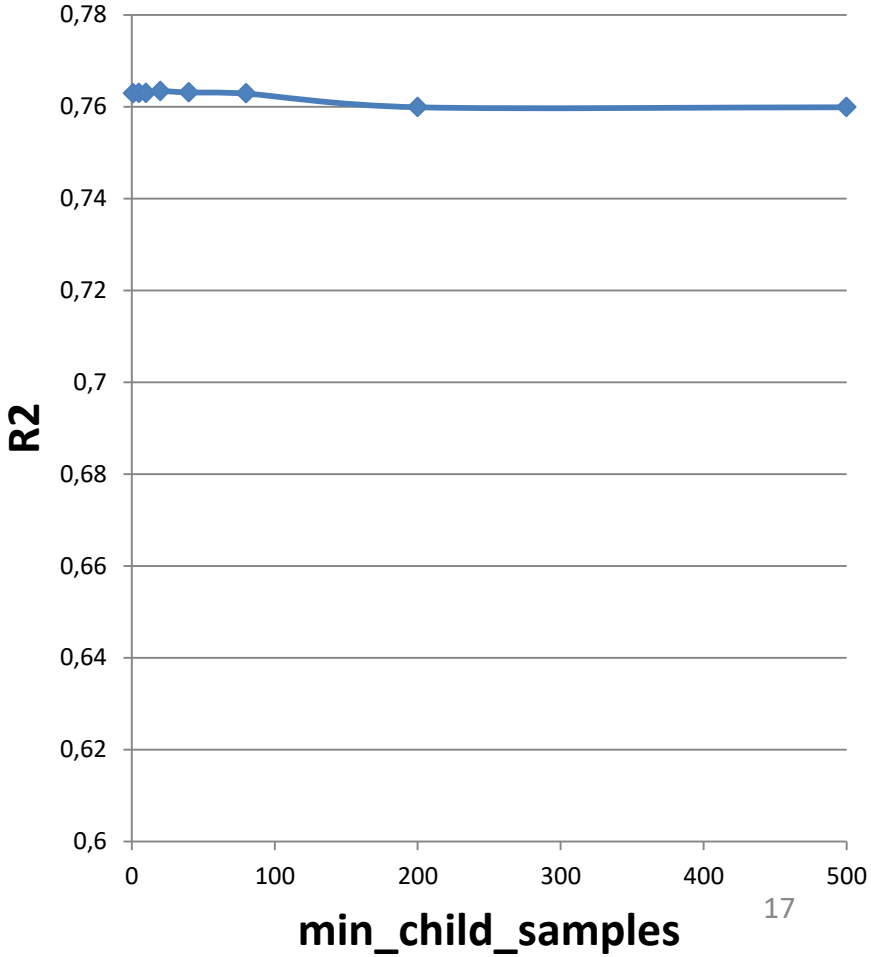


Настройка LGBM

Очень важно настроить параметр learning_rate.

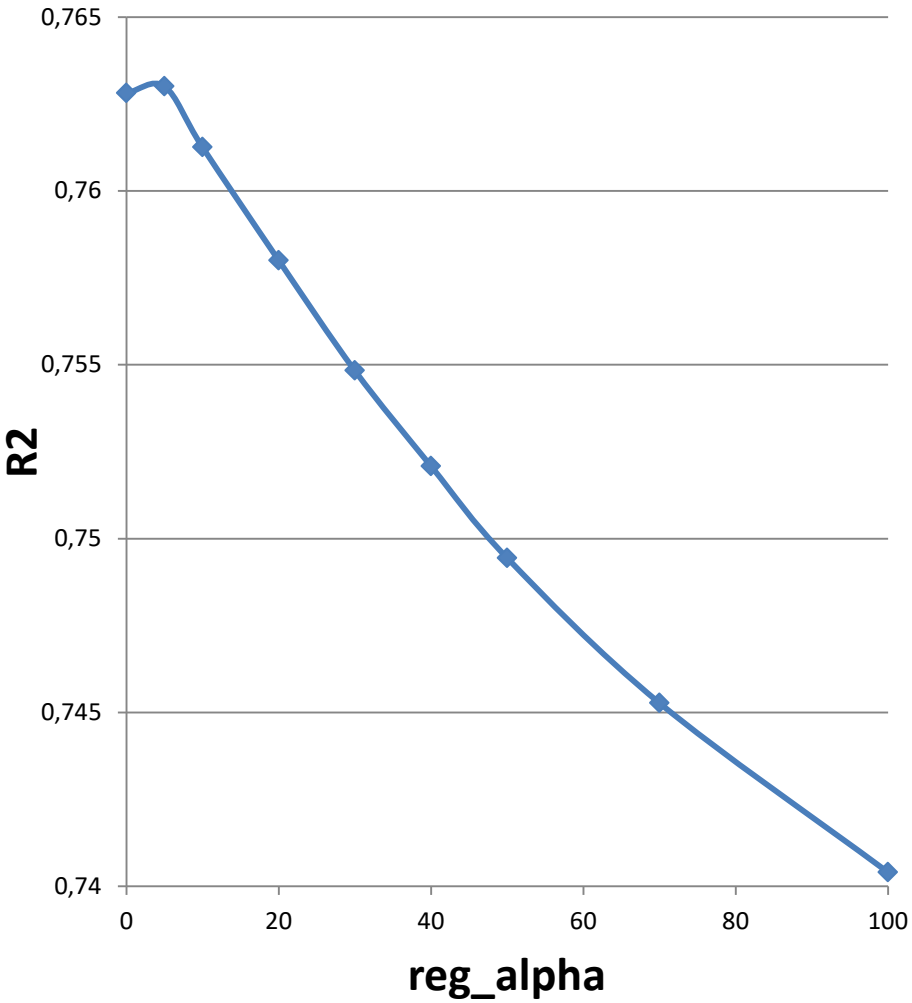
Остальные параметры играли очень небольшую роль.

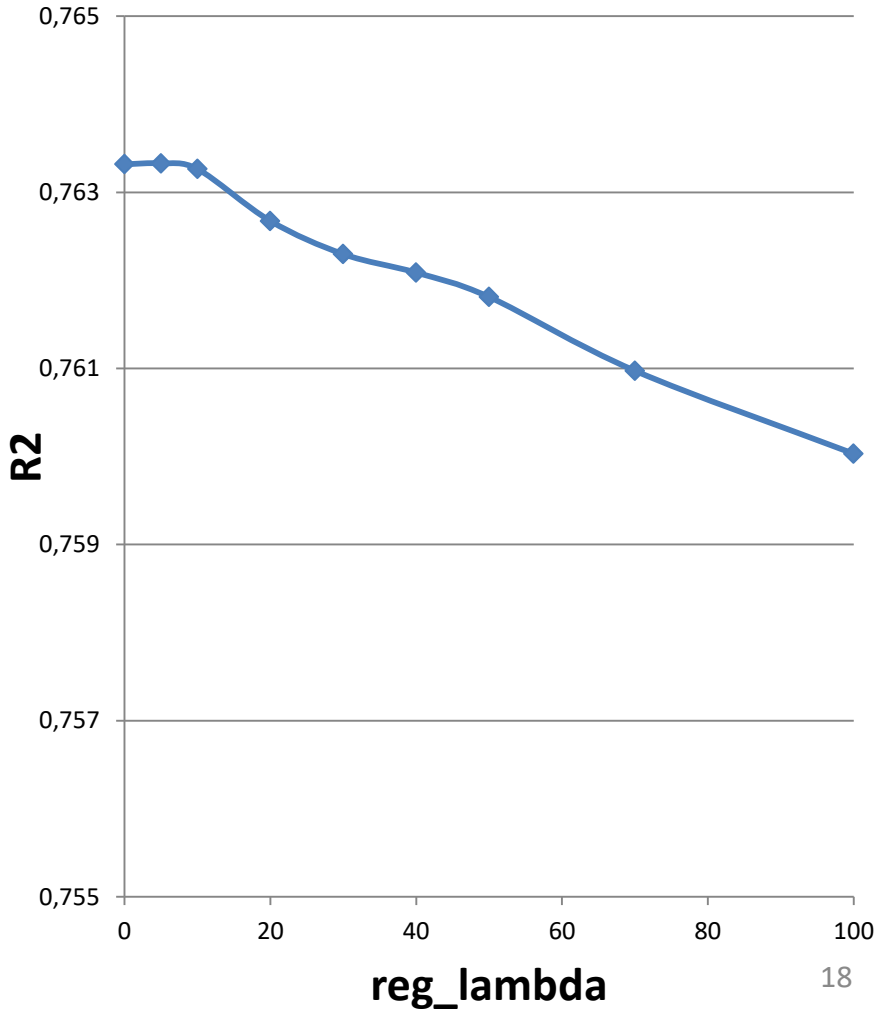




Настройка LGBM

Можно отметить параметры регуляризации (reg_alpha, reg_lambda), которые способны уменьшить переобучение LightGBM





Итоговая оценка и точность

Использование оптимального набора из 100 признаков, регрессии на основе LGBM и выборки из 25 000 000 элементов с валидацией на 3 000 000 элементах позволило получить финальный вариант оценки бонитета и возраста.

	R2	RMSE	MAE
Бонитет	0,867	0,503	0,37
Возраст	0,74	32,1	22,7

Как и ранее, погрешность возраста неоднородно распределена по разным породам.

Порода	**RMSE**	**MAE**		**Порода**	**RMSE**	**MAE**
Кедр	35,0	26,0		Береза	26,5	14,9
Сосна	34,6	25,1		Липа	24,3	15,0
Бук	31,0	22,8		Осина	23,0	13,1
Пихта	29,5	20,8				

Результаты

Для выбора признаков лучше всего работает метод добавления случайного шума.

Перестройка моделей с итеративным добавлением признаков работает лучше всего, но требует значительного объема вычислений.

Наилучшая регрессия была построена на основе градиентного бустинга в реализации LightGBM.

Использование большого набора признаков и объемной обучающей выборки позволяет достаточно точно оценить бонитет и возраст.