МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ

ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра Экспериментальной физики атмосферы

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

(бакалаврская работа)

На тему «Возможности прогнозирования на основе данных дорожных станций»

Исполнитель **Павленко Александр Викторович**

(фамилия, имя, отчество)

Руководитель кандидат физико-математических наук, доцент

(ученая степень, ученое звание)

Восканян Карина Левановна

(фамилия, имя, отчество)

«К защите допускаю»

Заведующий кафедрой

(подпись)

доктор физико-математических наук, профессор

(ученая степень, ученое звание)

Кузнецов Анатолий Дмитриевич

(фамилия, имя, отчество)

« » мая 2022 г.

Санкт–Петербург

2022

Содержание

1. Введение 3
1.1 Основная информация 4
1.2 Статистика 5
1.3 Состояния дорожного покрытия и причины их образования 10

2. Градиентный бустинг на решающих деревьях 15
3. Разработка модели предсказания состояния дорожного покрытия 21
3.1 Описание входных данных 21
3.2 Подробное описание параметров 22
4. Прогноз состояния дорожного покрытия 39
5. Заключение 41
6. Список использованных источников 44

Введение

Эксплуатация автомобильных дорог, как и любых других объектов и сооружений, находящихся под открытым небом, наиболее подвержена влиянию метеорологических условий. Поэтому при проектировании и использовании подобных объектов необходимо учитывать все особенности, с которыми можно столкнуться. Для этого Министерством транспорта создан ряд методических документов, на которые мы будем опираться в нашей работе.

Основной целью работы является выявление и демонстрация возможности прогнозирование состояния дорожного покрытия для дальнейшего принятия решений с целью снижения риска опасных ситуаций и дорожно-транспортных происшествий, а также увеличения срока службы объектов транспортной инфраструктуры.

Задачи:

- Ознакомиться с причинами возникновения аварий на автомобильных дорогах;
- Рассмотреть различные погодные явления, оказывающие влияние на состояние дорожного полотна и условия их образования;
- Изучить принцип функционирования модели машинного обучения градиентный бустинг на решающих деревьях;
- Собрать архивную базу дорожных метеорологических станций пригодную для выполнения цели работы;
- Определить возможное наличие в данных измерений пропусков и выбросов;
- Подготовить данные для корректной работы модели;
- Протестировать модель и оценить качество полученной информации.

1. Различные состояния дорожного покрытия, их описание, причины и опасности

1.1 Основная информация

Автомобильные дороги – сооружение, элемент транспортной инфраструктуры, предназначенный для передвижения транспортных средств, в том числе земельные участки в границах полос отвода автомобильных дорог и на них или под ними конструктивные элементы (дорожное покрытие, дорожное покрытие и подобные им элементы) и транспортное оборудование, являющиеся его технологическими частями, - защитные дорожные сооружения, искусственные дорожные сооружения, технологическое оборудование, элементы дорожных сооружений.

Отраслевые дорожные документы четко устанавливают определения для каждого вида опасных состояний дороги.

Согласно отраслевому методическому документу [1], виды отложений на поверхности дорожного полотна, различающиеся внешне:

- снежный накат
- рыхлый снег
- стекловидный лед

Перечислим признаки, по которым определяются соответствующие проблемные отложения:

Рыхлый снег образуется на поверхности автодороги ровным слоем одной толщины. Снег может быть мокрым, влажным и сухим, все зависит от содержания воды. Плотность свежевыпавшего снега может изменяться от 0,06 до 0,20 г/см^3. Из-за наличия рыхлого снега на дороге коэффициент сцепления может снижаться до 0.2.

Снежный накат – это слой снега, который был утрамбован и уплотнен автомобилями. Толщина варьируется от 5 до 80 миллиметров, а плотность от

0,3 до 0,6 г/см^3. Из-за наличия снежного наката коэффициент сцепления может снижаться до 0.1 - 0.25.

Стекловидный лед представляет собой тонкий слой льда (1 - 3 мм) и плотностью 0.7-1 г/см^3, напоминающего прозрачное стекло, а также в виде мутного неровного стекла толщиной более 10 мм, с плотностью от 0.5 до 0.8. Из-за наличия стекловидного льда коэффициент сцепления может снижаться до 0.08 - 0.15. Является самым опасным состоянием дорожного полотна.

Чтобы эффективно бороться с данными видами отложений необходимо грамотно учитывать все факторы, влияющие на их образование, усиление или ослабление, такие как метеорологические условия, географические и конструкционные особенности автодороги.

1.2 Статистика

Обледенение дорог, пожалуй, самая серьезная опасность, с которой сталкиваются жители, ежегодно приводящая к сотням серьезных травм и нескольким трагическим смертям.

Ежегодно в штате Вашингтон, США 24% дорожно-транспортных происшествий, связанных с погодными условиями, происходят на заснеженных, слякотных или обледенелых дорогах, а 15% — во время снегопада или мокрого снега [5]. Ежегодно более 1300 человек погибают и более 116 800 человек получают травмы в результате автомобильных аварий на заснеженном, слякотном или обледенелом тротуаре. Ежегодно около 900 человек гибнут и почти 76 000 человек получают травмы в автокатастрофах во время снегопада или мокрого снега. Снег и лед увеличивают расходы на содержание дорог. На содержание зимних дорог приходится примерно 20 процентов бюджета штата на содержание дорог. Государственные и местные агентства ежегодно тратят более 2,3 миллиарда долларов на операции по борьбе со снегом и льдом. Каждый год эти дорожные агентства также тратят миллионы долларов на устранение повреждений инфраструктуры,

вызванных снегом и льдом. В России статистика так же удручающая (табл. 1.1)

Таблица 1.1

Дорожно-транспортные происшествия, 2021 год

	ДТП		Погибло		Ранено		Тяжесть последствий
	абс	± % к АППГ	абс	± % к АППГ	абс	± % к АППГ	
Российская Федерация	**133331**	**-8.1**	**14874**	**-7.9**	**167856**	**-8.3**	**8.1**
Центральный федеральный округ	**33494**	**-5.7**	**3537**	**-8.2**	**41281**	**-5.1**	**7.9**
Северо-Западный федеральный округ	**14380**	**-9.9**	**1281**	**-7.6**	**17534**	**-11.1**	**6.8**
Южный федеральный округ	**14603**	**-6.7**	**1920**	**-8.9**	**18401**	**-7.1**	**9.4**
Северо-Кавказский федеральный округ	**5927**	**-6.5**	**1039**	**-10.9**	**7959**	**-7.5**	**11.5**
Приволжский федеральный округ	**29412**	**-8.0**	**3084**	**-6.4**	**37391**	**-8.6**	**7.6**
Уральский федеральный округ	**11276**	**-5.0**	**1159**	**-4.8**	**14733**	**-4.9**	**7.3**
Сибирский федеральный округ	**15527**	**-12.9**	**1802**	**-10.7**	**19634**	**-12.9**	**8.4**
Дальневосточный федеральный округ	**8659**	**-12.8**	**1049**	**-5.8**	**10861**	**-12.8**	**8.8**

где абс – абсолютное количество ДТП за данный период,
± % к АППГ – процентное изменение относительно аналогичного периода прошлого года

В России более 60% всех зимних дорожно-транспортных происшествий происходит в декабре [12]. Такое смещение в сторону первого месяца зимы происходит по нескольким причинам.

1. Несвоевременная замена типа резины с летней на зимнюю. Обычно жители нашей страны обуславливают эту проблему “внезапным приходом зимы”, но тем не менее, неподготовленность автомобиля к условиям езды в холодное время года по зимним дорогам оказывает значительное влияние аварийность на дорогах. Помимо проблем с резиной можно отметить и другие технические неисправности автомобилей, такие как неработающие электронные системы безопасности (ABS), износ тормозной системы и подвески, недостаточные условия видимости из-за неисправных боковых зеркал или лобового стекла.

2. Вторым немаловажным фактором является неготовность самих водителей к езде по зимним дорогам. Езда зимой требует от водителя большей внимательности, аккуратности и готовности к быстрому принятию решений, и у некоторых водителей адаптация к этим условиям происходит медленнее, чем этого требует ситуация.

Наибольшее количество ДТП (рис. 1.1) происходит в субъектах Центрального федерального округа – Московская и Тверская области, а также Северо-Западного федерального округа – Ленинградская область, а также в Краснодарском крае и Ростовской области, в Башкирии, Мордовии, в Пермском крае, Самарской и Оренбургской областях.

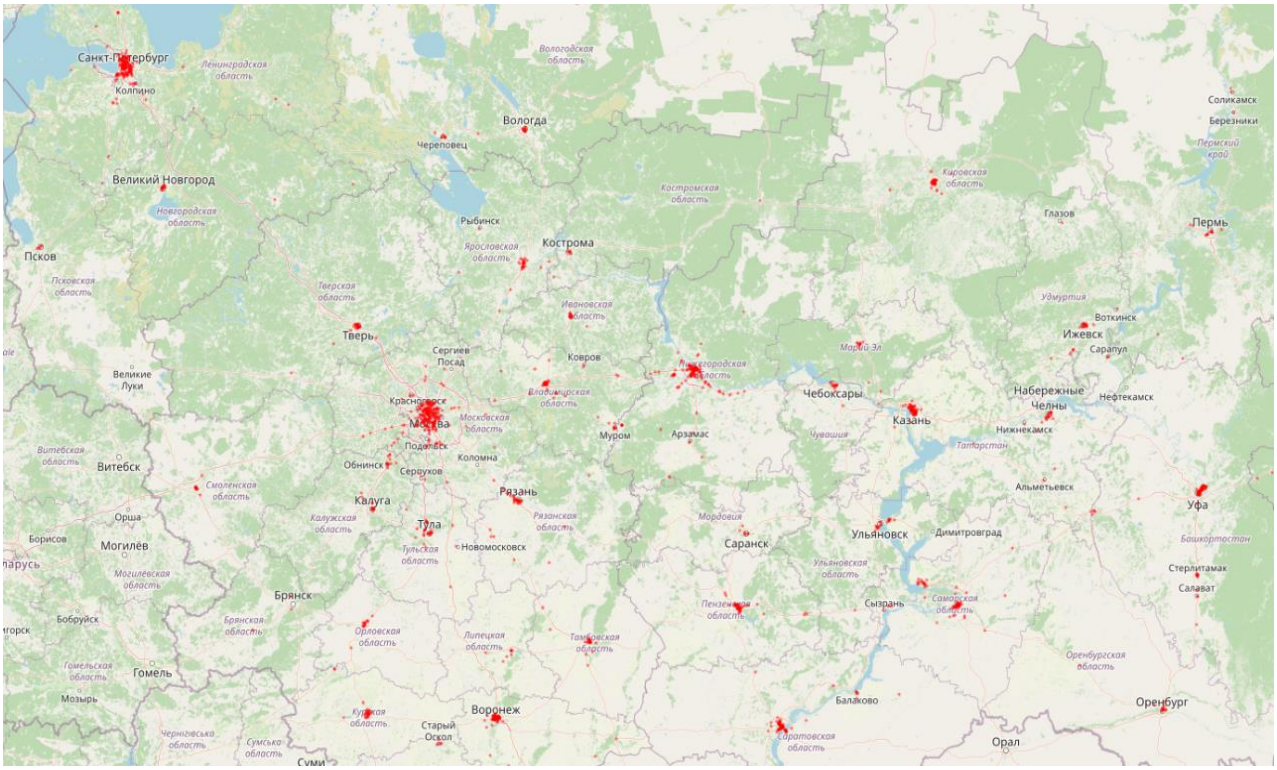


Рисунок 1.1 – Места концентрации ДТП в европейской части России в декабре 2021

Места концентрации ДТП в Санкт-Петербурге в декабре 2021 показаны более подробно на рисунке 1.2.

Рисунок 1.2 – Места концентрации ДТП в Санкт-Петербурге в декабре 2021

Посмотрим на распределение количества ДТП в Санкт-Петербурге и области. Исходя из данных рисунка 1.3, можно извлечь ключевое положение [12].

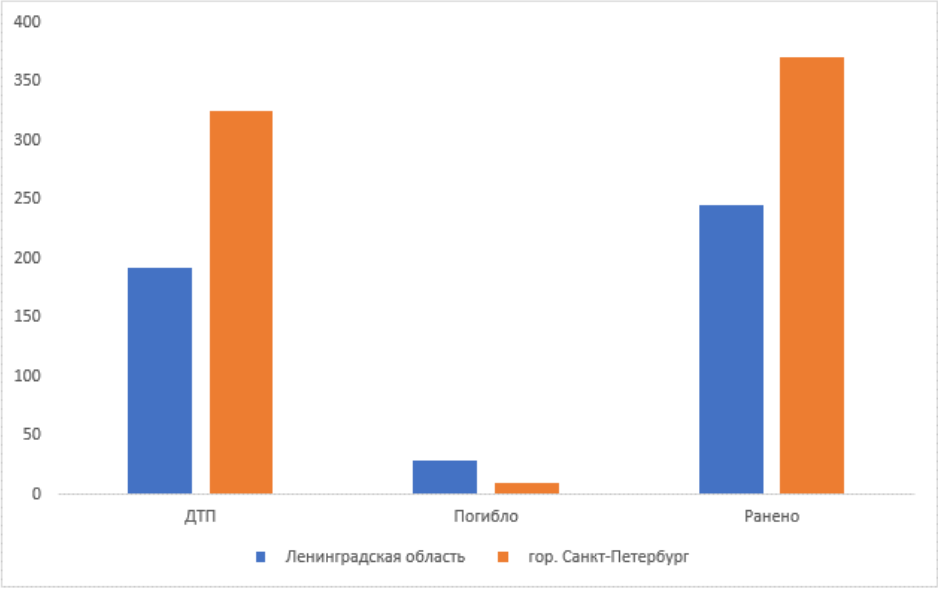


Рисунок 1.3 – ДТП в Санкт-Петербурге и Ленинградской области

В Санкт-Петербурге наблюдается большее количество ДТП с большим количеством раненых, но в Ленинградской области количество смертей превышает количество смертей в Санкт-Петербурге (28 и 9 соответственно) [12].

Это происходит по нескольким причинам:

1. Скоростной режим в городе. В городе средняя скорость автомобильного движения существенно меньше, чем в области.
2. Скорость прибытия аварийных служб в городе существенно быстрее, чем в области, что позволяет своевременно оказать медицинскую помощь пострадавшим.
3. Состояние дорожного покрытия. В городе на очистку и обработку дорог затрачивается существенно больше ресурсов, чем на соответствующие мероприятия в области.

На примере статистических данных (табл. 1.2), собранным в штате Айдахо, США, покажем, что среднемесячное количество ДТП уменьшается более чем на 80% благодаря использованию противообледенительной обработки [4].

Таблица 1.2

Показатели эффективности зимнего технического обслуживания в штате Айдахо, США

	с 1992 по 1997 (Без противообледенительной обработки)	с 1997 по 2000 (С противообледенительной обработкой)	Процент уменьшения
Количество ДТП	16.2	2.7	83%

Таким образом, можно сделать вывод о том, что поддержание автомобильных дорог в надлежащем состоянии является очень важной задачей, от своевременного и качественного решения которой напрямую зависит безопасность движения транспорта и человеческие жизни.

И для оптимизации расходов для решения данной задачи можно использовать метод прогнозирования состояния дорожного покрытия, который мы рассмотрим в практической части работы.

1.3. Состояния дорожного покрытия и причины их образования

Рыхлый снег. Образование на дорожном покрытии происходит при выпадении снега и других видов твердых осадков, при температуре воздуха от -6°до -10°С снег не уплотняется при относительной влажности воздуха менее 90%, а сохранение его происходи при низких температурах, так как процедура уплотнения снега автомобилями замедляется.

Снежный накат. Образование происходит при наличии влажного снега на дорожном покрытии под действием автомобильного транспорта и определенных метеорологических условиях. Наибольшая вероятность образования снежного наката происходит при следующих погодных условиях:

- ✓ осадки в виде снега при температуре воздуха от -0 ° до - 6 °С;
- ✓ при температуре воздуха от -6 ° до - 10 °С образование снежного наката происходит при влажности воздуха выше 90 %;
- ✓ при температуре больше 0°С образование происходит при высокой интенсивности осадков (снега), когда таяние снега происходит медленнее, чем его оседание.

Причинами образования льда на дороге являются отрицательные температуры и влажность (вода) на поверхности, комбинация, которая может возникать несколькими способами:

- Мороз

- Туман, проходящий над холодной поверхностью проезжей части
- Замерзание просачивающихся грунтовых вод или талого снега
- Намерзание снега, который первоначально растаял на теплом дорожном покрытии.
- Ледяной дождь.

Морозы обычно случаются в холодные, относительно ясные ночи, когда скорость ветра низкая. В ясную ночь поверхность испускает инфракрасное излучение, и без облаков, останавливающих его, большая часть этого излучения теряется в космосе. Поэтому поверхность и воздух вблизи нее быстро охлаждаются. Вдали от холодной поверхности воздух имеет тенденцию быть более теплым.

В пасмурные ночи облака замедляют или предотвращают потерю тепла с поверхности, и поэтому охлаждение происходит намного меньше. Сильные ветры также, как правило, препятствуют охлаждению поверхности, поскольку ветреные условия «возмущают» атмосферу и смешивают часть более теплого воздуха наверху с землей. Если температура проезжей части и температура точки росы выше точки замерзания, на поверхности образуется жидкая вода (роса), а если температура ниже нуля, вместо нее образуется иней (рис. 1.4).

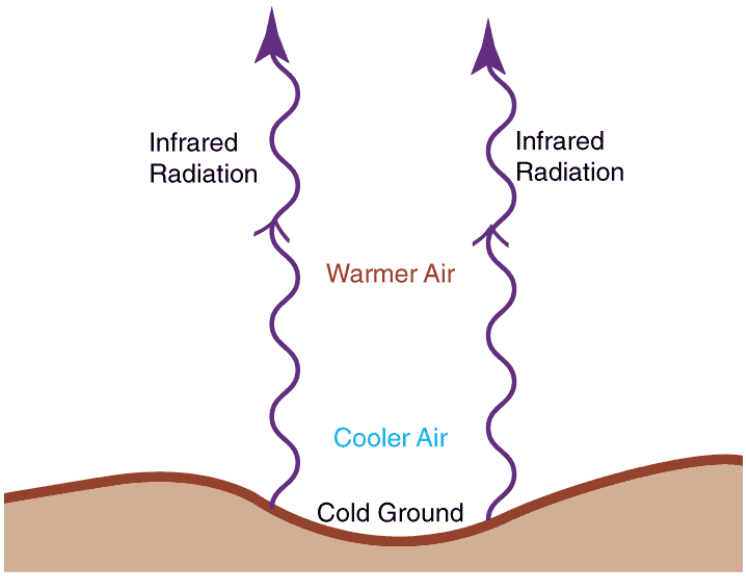


Рисунок 1.4 – Радиационное выхолаживание поверхности

Влажная атмосфера с большим количеством водяного пара способствует замерзанию, поскольку она может дать больше воды для замерзания. Такие влажные условия часто возникают после периода осадков или во влажных местах, например, вблизи заболоченных участков или рек. Обилие водяного пара также повышает температуру точки росы, поскольку с большим количеством водяного пара в воздухе не нужно так сильно охлаждать воздух, чтобы получить росу или иней. Как отмечалось выше, в ветреные периоды меньше мороза, поскольку перемешивание, вызванное ветром, приносит более теплый и сухой воздух сверху вниз к поверхности. Мороз часто более распространен в долинах и низинах, куда холодный воздух имеет тенденцию стекать и скапливаться (подробнее об этом ниже) и где скорость ветра обычно меньше.

Иней обычно накапливается медленно и редко накапливается более чем на 15 мм. По этой причине мороз может представлять меньшую угрозу, чем туман и другие формы обледенения.

Таким образом, заморозки обычно случаются относительно ясными ночами при слабом ветре. Сильные ветры препятствуют образованию инея. Большинство метеостанций сообщают о температуре на высоте около 2 м, где температура может быть значительно выше, чем на поверхности. Так что, если на небе относительно безоблачно, а ночная температура воздуха падает до 30 градусов, заморозки могут представлять реальную угрозу.

Хотя риск обледенения из-за мороза уменьшается из-за медленного накопления и относительно минимальной толщины, это не относится к обледенению дороги, связанному с туманами.

Туман часто образуется в холодные, ясные ночи, когда температура падает до точки росы. Туман содержит большое количество жидкой воды, и если над дорогой, остывшей до температуры ниже точки замерзания,

проходит полоса тумана, обледенение может быть быстрым и сильным, а толстый слой льда нарастает в течение нескольких минут.

Типичный сценарий начинается с ясной холодной ночи, когда поверхность быстро остывает. На дороге может образоваться небольшое обледенение, а на мокрых поверхностях может начаться туман. Туман распространяется, и толстый слой льда ложится на дорогу.

Опасные условия обледенения часто возникают, когда температура дорожного покрытия днем выше нуля, а ночью опускается ниже нуля градусов, т.е. колеблется около нулевой отметки. Даже если дорога свободна от снега, снег часто встречается по ее сторонам. Особенно это касается дорог, которые активно расчищают, с завалами снега, примыкающими к открытым полосам. Придорожный снег тает в течение дня, особенно на участках, прилегающих к относительно теплой дороге (дороги, особенно с темным асфальтовым покрытием, легко поглощают солнечное тепло). Талая вода течет по дороге днем, а ночью замерзает, особенно если на небе мало облаков.

Аналогичная ситуация может возникнуть и без снега, если на дорогу стекает вода из родника или другого источника воды. Днем вода остается жидкой, а ночью замерзает на дорожном полотне. Кроме того, мокрые дороги часто очень быстро замерзают, когда воздух над ними сухой. Причиной этого является охлаждение от испарения. Все мы испытываем холод от испарительного охлаждения, когда выходим из душа или ванны — это охлаждение сильнее всего, когда воздух вокруг нас сухой. Таким образом, в холодную сухую ночь испарение с мокрой дороги может привести к тому, что поверхность охлаждается намного быстрее, чем сухие дорожные покрытия поблизости, что приводит к локальному обледенению. Таким образом, по всем этим причинам важно, чтобы персонал дорожного хозяйства был знаком с влажными участками дорог и часто проверял их холодными ночами, когда температура воздуха падает до минус 30°C.

Особенно опасный вид обледенения возникает в начале зимнего сезона или после периода теплой погоды. В такие моменты дорожное покрытие промерзает. Если погода становится холодной, может начать падать снег, который сначала растает в мокрую слякоть на теплом дорожном покрытии. Если температура воздуха продолжает быстро падать (возможно, после прохождения арктического фронта с севера или проталкивания холодного), прогревающееся от поверхности дороги и теплой земли внизу перемешивается с холодным воздухом, и смесь слякоти превращается в лед.

Ледяной дождь возникает, когда у поверхности есть слой воздуха с температурой ниже точки замерзания и более теплый (выше точки замерзания) воздух наверху. Дождь с высоты попадает в холодный слой и охлаждается до минусовой температуры – и остается жидким. Когда такой переохлажденный дождь попадает на поверхность, он тут же замерзает в прозрачную гололедицу. Такой ледяной дождь часто приводит к ледяным бурям, которые могут сделать путешествие опасным и привести к обрушиванию деревьев и линий электропередач.

В данных из отраслевого дорожного методического документа (табл. 1.3) кратко сформированы основные условия образования ключевых состояний обледенения.

Таблица 1.3

Классификация различных видов зимней скользкости дорожных покрытий и условия их образования

Вид зимней скользкости	Условия образования				
	Температура воздуха	Температура покрытия	Осадки, их вид	Состояние покрытия	Дополнительные условия
Гололедица	Ниже 0°С	Ниже 0°С	Любые, выпадающие при температуре воздуха выше -3°С	Мокрое	Выпадение осадков предшествует образованию скользкости
	Выше 0°С	Ниже 0°С	Жидкие	-	-
	От 0°С до -5°С	Ниже 0°С	Мокрый снег	-	Количество осадков, зафиксированное метеостанцией (Q=0 мм)
"Черный лед"	То же	Ниже 0°С, ниже точки росы	Нет	Сухое	-
Гололед	Ниже 0°С	Ниже 0°С	Переохлажденные жидкие (дождь, морось)	Любое	-
Снежный накат	От 2°С до 0°С	-	Твердые (снег, мокрый снег)	-	-
	От 0°С до -6°С	-	То же	-	Интенсивность снегопада не менее 0,6 мм/ч
	От - 6°С до -10°С	-	То же	-	Относительная влажность воздуха не менее 90%

2 Градиентный бустинг на решающих деревьях

В данной работе основной моделью машинного обучения является градиентный бустинг на решающих деревьях. Начнем с деревьев.

Решающее дерево представляет собой набор решений (правил), по которым производится решение о спуске на следующий лист. Принцип действия напоминает процесс принятия решения человеком. Легче всего понять процесс на простом примере (табл. 2.1).

Таблица 2.1.

Пример выборки данных

	x1	x2	y
n1	7	2	red
n2	1	9	blue
n3	4	1	yellow
n4	9	7	black
n5	6	4	?

Пусть у нас имеются некоторые данные n1, n2, n3, n4, n5 с некими параметрами x1 и x2, а также целевым параметром y.

Составим решающее дерево на основе данного набора (рис. 2.1):

Таким образом, можем наблюдать список составленных правил, по которым определяются значения целевой переменной y. Давайте, на примере элемента n5 определим его принадлежность к классу.

Значение x1 > 5, поэтому мы пойдем по левой ветви. x2 < 5, поэтому мы можем сделать вывод о значении целевой переменной – red.

Понять принцип построения решающих деревьев можно также взглянув на график принадлежности классов (рис. 2.2).

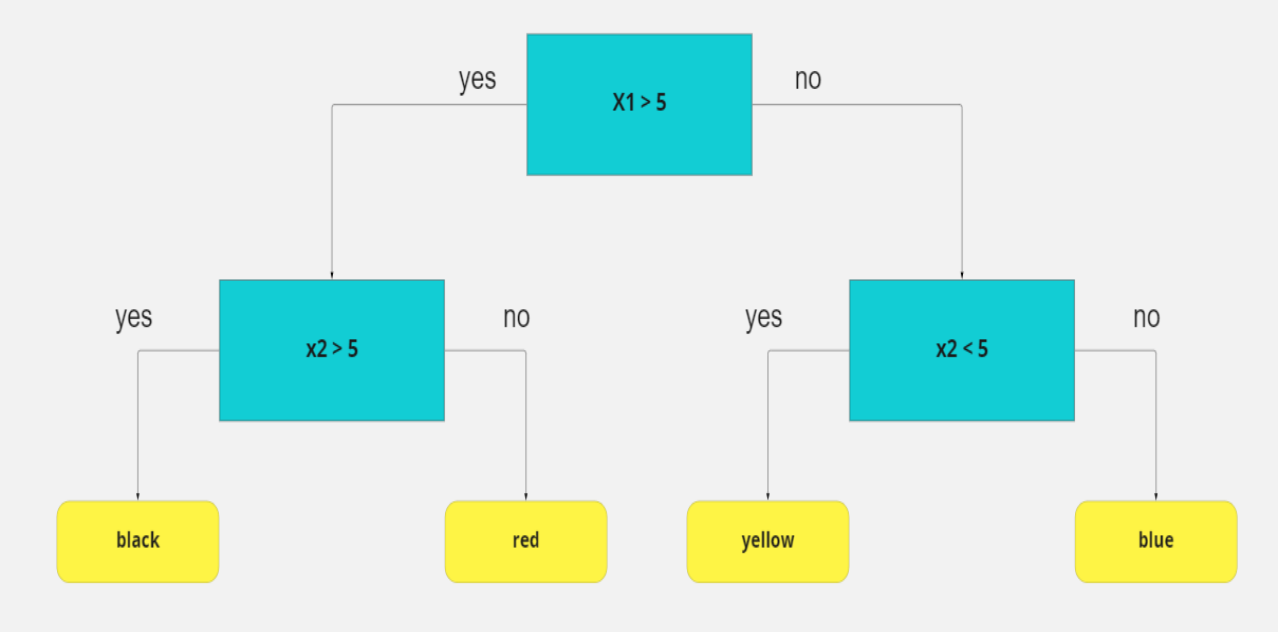


Рисунок 2.1 – Схема решающего дерева.

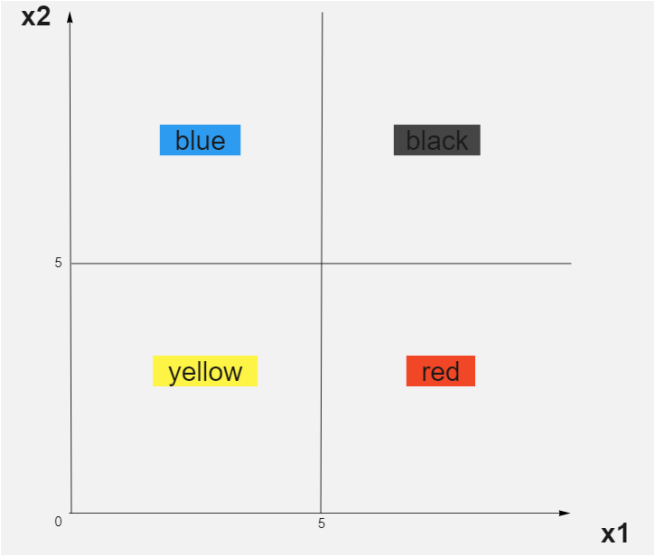


Рисунок 2.2 – Разделяющие поверхности классов

Каждое правило порождает разделение текущего подмножества пространства на 2 части. Таким образом, любому объекту в области определения можно присвоить соответствующий класс.

Стоит отметить, что в узлах принятия решений могут находиться не только простые условия (x>const), но и сложные функции, комбинации признаков и их производные.

Но не стоит забывать об оверфиттинге, или же о проблеме переобучения.

Дело в том, что мы можем построить сколь угодно большой алгоритм, или ансамбль моделей, который будет иметь 100% точность на обучающей выборке. Но когда на вход данного алгоритма попадут данные, которые не были использованы при обучении, его точность существенно упадет. Это происходит из-за того, что алгоритм не смог найти общих закономерностей в выборке, а использует только локальные, представленные в обучающей выборке. Проще всего это можно понять, используя графические разделяющие поверхности (рис. 2.3):

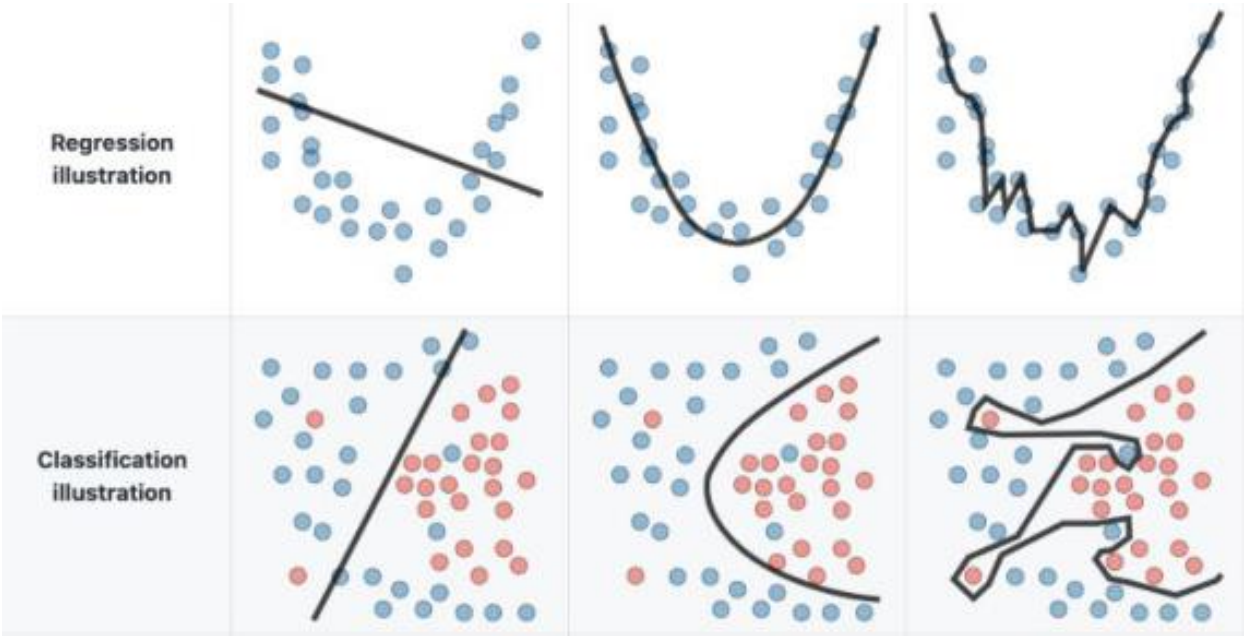


Рисунок 2.3 – Демонстрация переобучения в задачах регрессии и классификации

Как мы видим, разделяющая поверхность в задачах классификации учитывает все отклонения точек от оптимальной разделяющей поверхности, характерных только для данной выборки, и поэтому качество на тестовой выборке будет существенно хуже. Оптимальная же разделяющая поверхность допускает наличие ошибочных предсказаний, но при этом описывает общие закономерности в данных, что позволяет применять модель на генеральной совокупности.

Бороться с данной проблемой можно несколькими способами. Например, ограничивать глубину дерева (обрезать листья), то есть установить предельно допустимое количество сплитов по правилам в глубину (рис. 2.4).

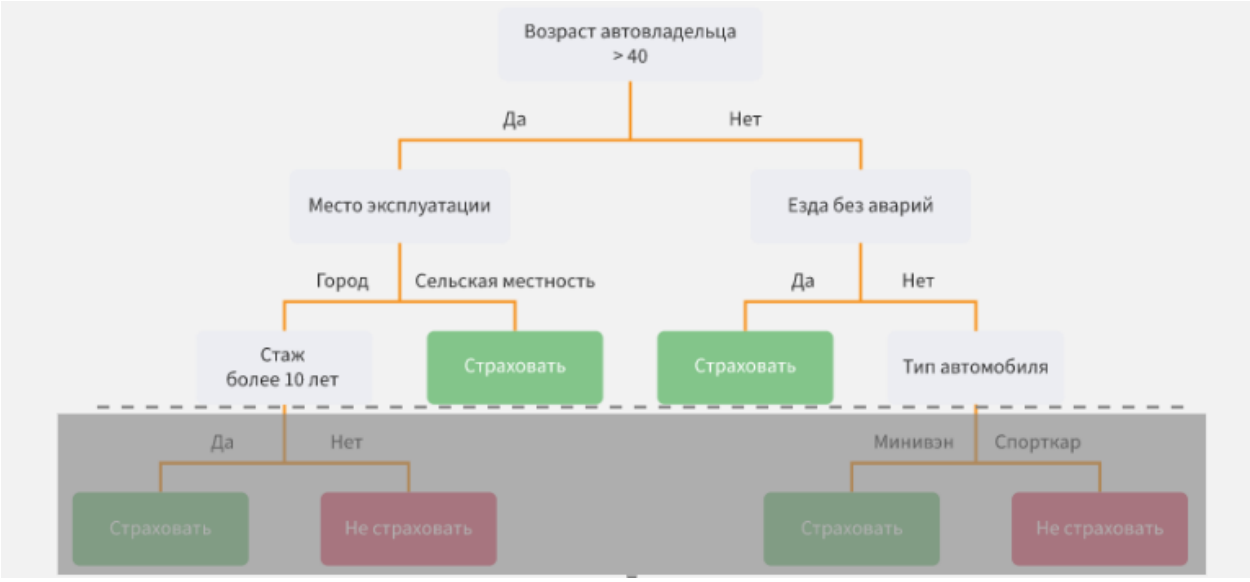


Рисунок 2.4 – Ограничение глубины дерева.

Например, возьмем решающее дерево глубиной 3, и ограничим его глубиной 2. Таким образом, мы будем иметь 2 определенных листа, и 2 листа с неопределенностью. Устранить ее можно разными способами, например, превратить предиктат в лист путем сравнения количества экземпляров классов, попадающих в эту ветвь (рис. 2.5).

Рисунок 2.5 – Ограничение глубины дерева. Преобразование листьев

Итого имеем соотношения 65% на 35% и 83% на 27%, поэтому данные предиктаты преобразуются в листья с классом “Не страховать”.

Другим способом является использование критерия останова. Применяя данный метод, разделение ветвей будет останавливаться, если достигнуто выбранное условие, например точность, информативность, или же вышеупомянутое соотношение классов.

Решающие деревья можно объединять, получая ансамбли моделей (леса).

Делать это можно также несколькими способами, один из которых – беггигнг.

Объяснить принцип действия беггинга можно на примере мешка. Каждый раз мы берем не всю обучающую выборку, а лишь ее часть, например 60% или 75%, таким образом, каждый раз мы будем получать небольшую обучающую выборку, отличную от основной, увеличивая универсальность модели и уменьшая привязку к начальной выборке.

Помимо выбора случайной части экземпляров класса, для разных деревьев мы можем также выбирать и разные наборы фичей (аргументов)

(рис. 2.6), что еще сильнее увеличивает универсальность модели. Стоит отметить, что использовать беггинг фич следует использовать при действительно большом наборе (более 15-20).

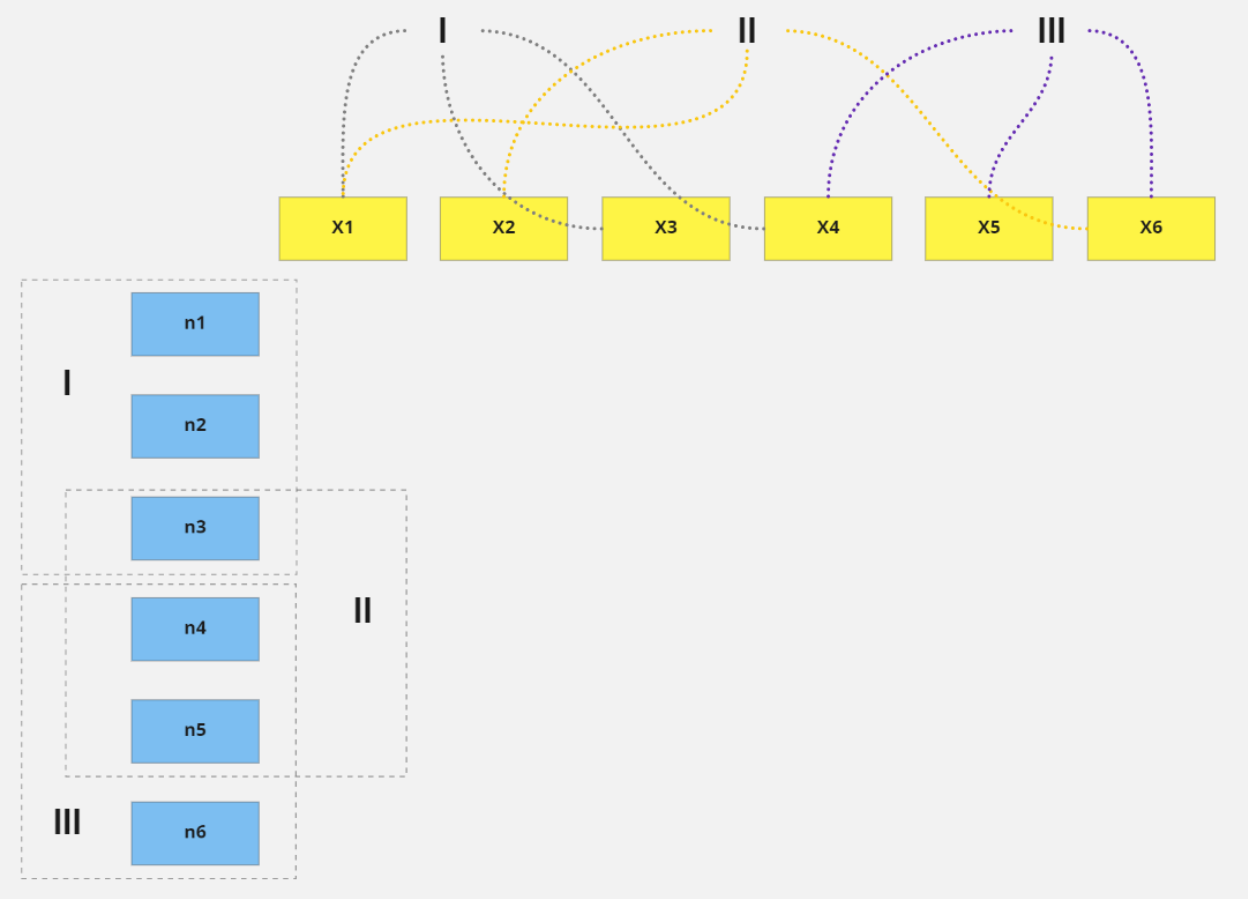


Рисунок 2.6 – Принципы беггинга

3 Разработка модели предсказания состояния дорожного покрытия

3.1 Описание входных данных

Исследование проводится с использованием данных, представленных в приложении к работе Grabowski, Dariusz, 2020, "Road CCTV images with asso ciated weather data", размещенной в открытом доступе на сайте.[10]

Набор данных представляет собой таблицу, в которой зафиксированы измерения с автоматических метеостанций, расположенных на автомобильных дорогах на территории Польши. Измерения проводились в период с ноября 2018 года по март 2019 года и содержат информацию о показателях:

- температура, измеренная на различных высотах (0, 20 см, 2 м);
- температура точки росы;
- относительная влажность;
- тип осадков;
- состояние дорожного покрытия;
- скорость и направление ветра;

В наборе содержится более 3.3 миллиона записей данных с 551 станции, разделенные по времени с промежутком в 20 минут (табл. 3.1).

Таблица 3.1

Общий вид набора данных

						temperature					
stationId	meteoDate	dew	humidity	precipitation	roadCondition	0cm	20cm	2m	warnings	wind direction	wind speed[m/s]
10-0	2019-03-12 23:00:00	0	94.6	none	dry	-0.5	0	-1.3	none	221	0
10-0	2019-03-12 23:20:00	0	99.2	none	dry	-0.7	0	-1.4	none	218	0
10-0	2019-03-12 23:40:00	0	99.2	none	dry	-0.9	0	-1.4	none	195	0
10-0	2019-03-13 00:00:00	0	99.2	none	dry	-1	0	-1.7	none	203	0
10-0	2019-03-13 00:20:00	0	99.2	none	dry	-1	0	-1.7	none	207	0

3.2 Подробное описание параметров

3.2.1 Состояние дорожного покрытия

Начнем рассматривать набор данных с целевой переменной – состояния дорожного покрытия (roadCondition). Всего в данной колонке имеются 8 уникальных значений переменной (рис. 3.1):

- dry – сухая поверхность дорожного полотна
- wet – влажная поверхность дорожного полотна
- moist - влажная поверхность дорожного полотна (но менее влажная, чем ‘wet’)
- saline – обработанная поверхность дорожного полотна
- snow – покрытая снегом поверхность дорожного полотна
- rimeice – покрытая изморозью поверхность дорожного полотна
- ice – покрытая льдом поверхность дорожного полотна
- unknown – состояние поверхности дорожного полотна неизвестно

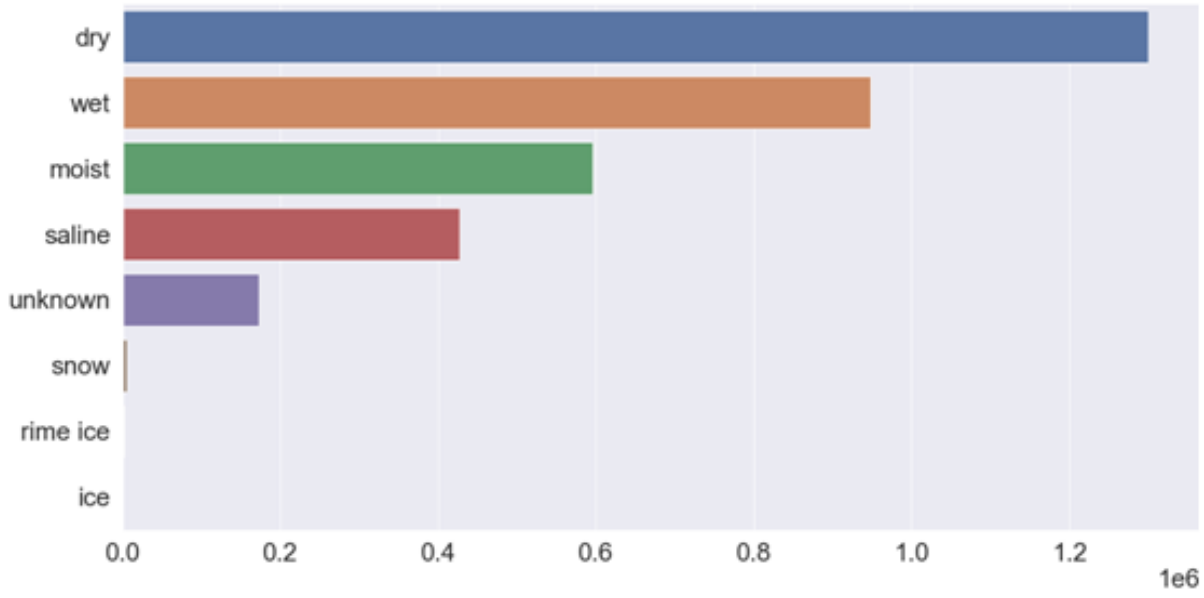


Рисунок 3.1 – Диаграмма распределения классов

Рассматривая диаграмму распределения классов в наборе данных, мы можем наблюдать сильный дисбаланс классов:

- dry – 1300120 экземпляров
- wet – 948322
- moist – 595614
- unknown – 171960
- saline – 426672
- snow – 4977
- rime ice – 1893
- ice - 121

Количество экземпляров, желаемых для обнаружения (опасные состояния: snow, rimeice, ice) на порядок меньше других состояний. В дальнейшем это может серьезно повлиять на качество модели.

Решить проблему дисбаланса можно несколькими способами, но в нашем случае будем использовать один из простейших – Oversampling. Суть его заключается в том, что мы увеличиваем количество экземпляров класса до необходимого нам уровня дублированием уже существующих экземпляров (рис. 3.2, табл. 3.2).

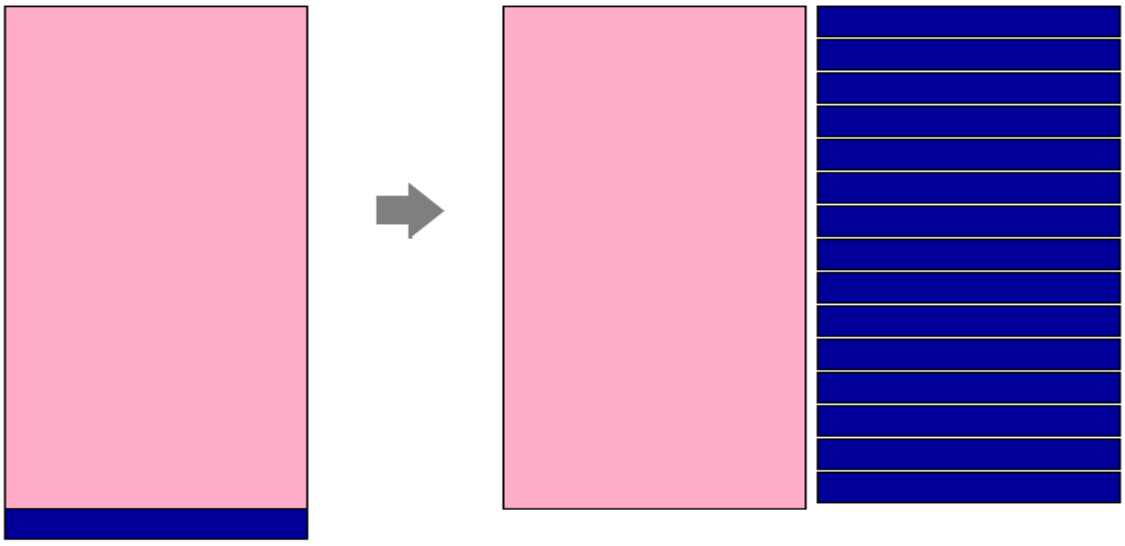
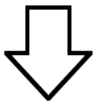


Рисунок3.2 – Демонстрация Oversampling’а

Таблица 3.2

Демонстрация Oversampling’а

						temperature					
stationId	**meteoDate**	**dew**	**humidity**	**precipitation**	**roadCondition**	**0cm**	**20cm**	**2m**	**warnings**	**wind direction**	**wind speed[m/s]**
10-0	2019-03-13 00:40:00	0	99.3	none	dry	-1.2	0	-1.7	none	194	0
100-0	2018-11-18 04:16:55	-6	91.4	unknown	dry	-3	0	-4.9	high danger	246	1.4
10-0	2019-03-13 05:00:00	0	99.2	none	dry	0.1	0	-0.1	none	173	0
100-0	2018-11-18 07:06:34	-5.7	92.1	unknown	moist	-3.3	0	-4.7	icing danger	284	1.5
1074-0	2018-12-27 07:10:00	4	99.4	none	wet	3.1	0	4.1	none	224	1.9
VP1201-0	2018-12-25 03:30:00	-3.4	93	snow	ice	-2	0	-2.4	icing danger	267	0



						temperature					
stationId	**meteoDate**	**dew**	**humidity**	**precipitation**	**roadCondition**	**0cm**	**20cm**	**2m**	**warnings**	**wind direction**	**wind speed[m/s]**
10-0	2019-03-13 00:40:00	0	99.3	none	dry	-1.2	0	-1.7	none	194	0
100-0	2018-11-18 04:16:55	-6	91.4	unknown	dry	-3	0	-4.9	high danger	246	1.4
10-0	2019-03-13 05:00:00	0	99.2	none	dry	0.1	0	-0.1	none	173	0
100-0	2018-11-18 07:06:34	-5.7	92.1	unknown	moist	-3.3	0	-4.7	icing danger	284	1.5
100-0	**2018-11-18 07:06:34**	**-5.7**	**92.1**	**unknown**	**moist**	**-3.3**	**0**	**-4.7**	**icing danger**	**284**	**1.5**
1074-0	2018-12-27 07:10:00	4	99.4	none	wet	3.1	0	4.1	none	224	1.9
1074-0	**2018-12-27 07:10:00**	**4**	**99.4**	**none**	**wet**	**3.1**	**0**	**4.1**	**none**	**224**	**1.9**
VP1201-0	2018-12-25 03:30:00	-3.4	93	snow	ice	-2	0	-2.4	icing danger	267	0
VP1201-0	**2018-12-25 03:30:00**	**-3.4**	**93**	**snow**	**ice**	**-2**	**0**	**-2.4**	**icing danger**	**267**	**0**

Но мы не можем начать процедуру oversampling’а без предварительной ‘очистки’ нашего набора данных. Суть очистки заключается в удалении или заполнении нулевых значений, удалении заведомо неверных значений, ограничении крайних значений, восстановлении неизвестных значений.

Начнем рассматривать остальные параметры по порядку.

3.2.2 Значения температуры точки росы

Рассматривая гистограмму распределения значений температуры точки росы (рис. 3.3), можно заметить, что в наборе данных имеется очень большое количество нулевых значений. Таким образом, можно представить 2 гипотезы:

1. это действительные значения, показывающие реальные данные о температуре точки росы на станциях;
2. это пропущенные значения, то есть неизвестное значение в таблице заменяется на нулевое.

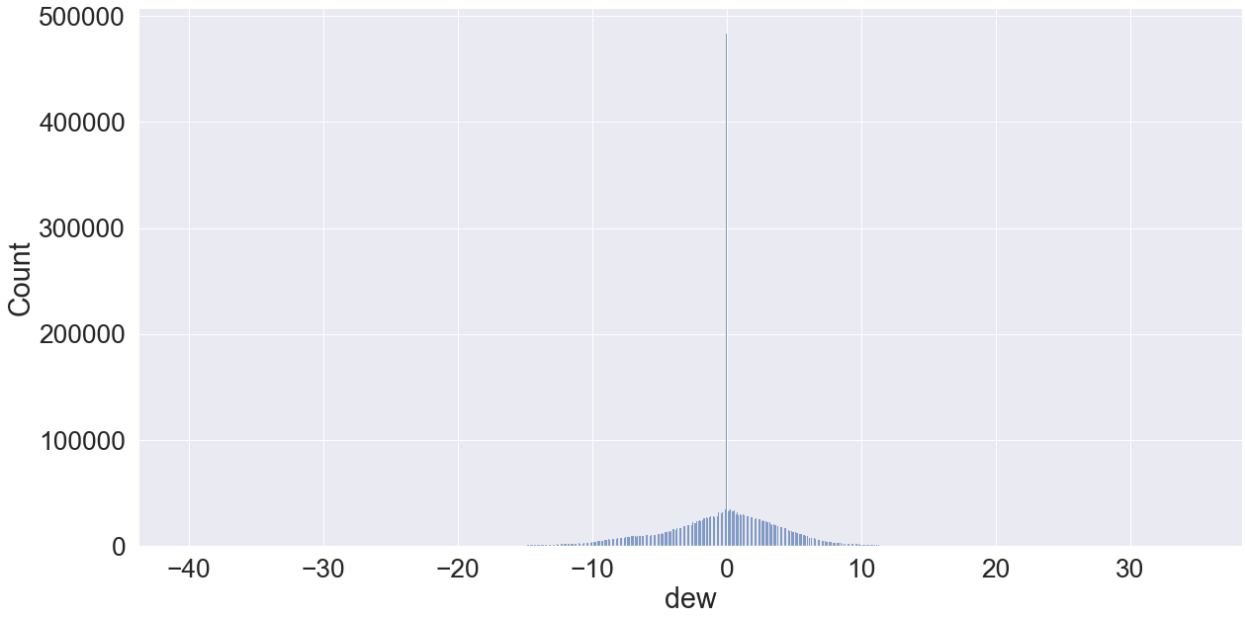


Рисунок 3.3 – Гистограмма распределения значений температуры точки росы

При этом, если удалить из набора данных нулевые значения температуры точки росы, то мы получим более адекватную гистограмму, с распределением, близким к нормальному (рис. 3.4).

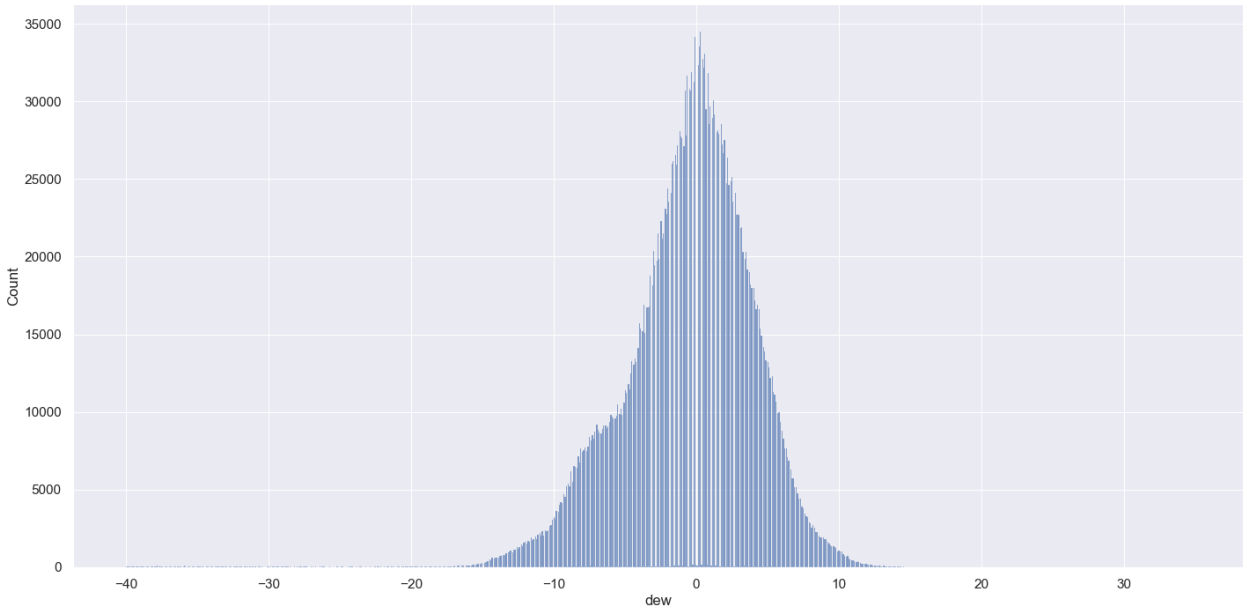


Рисунок 3.4 – Гистограмма распределения значений температуры точки росы с удаленными нулевыми значениями

Таким образом, будем считать, что нулевые значения были получены из отсутствующих данных. Применим к строкам с этими значениями пониженные веса.

Также в нашем наборе данных присутствуютнекоторые значения, сильно отклоняющиеся от мер центральной тенденции. Взглянем на крайние значения нашей выборки:

- min = -40 °C
- max = 34.6 °C

Такие значения могут сильно повлиять на качество нашей модели, поэтому мы обрежем их по значению 0.05 и 0.95 квантилей нашей выборки.

3.2.3Относительная влажность

Рассмотрим гистограмму распределения значений относительной влажности из нашего набора данных (рис. 3.5).

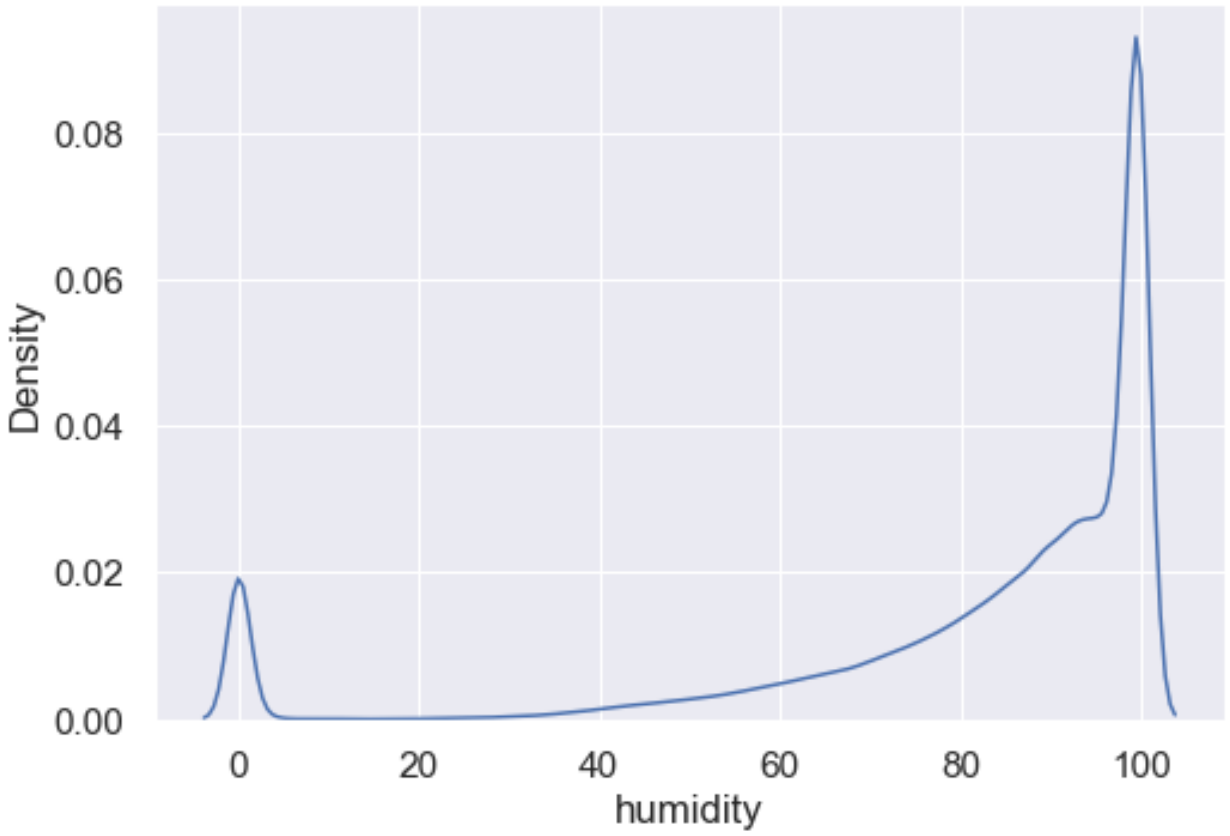


Рисунок 3.5 – График плотности распределения значений относительной влажности

На графике нетрудно заметить пик нулевых значений влажности. Зная о том, что пропущенные значения температуры точки росы заменялись нулевыми значениями, будем считать, что с относительной влажностью поступали так же.

Можно попытаться восстановить эти данные, используя значения температуры точки росы и температуры на высоте 2 м используя формулу:

$$rh = \frac{e(T_d)}{e_s(T)} \tag{1}$$

где

$e(T_d)$ - давление насыщенного пара для температуры точки росы

$e_s(T)$ - давление насыщенного пара для температуры

Используя ненулевые данные относительной влажности, мы можем определить погрешность данного метода восстановления. Полученные результаты (табл. 3.3) представлены в виде расчетной влажности и фактической влажности для более удобного сравнения.

Таблица 3.3

Расчетные значения относительной влажности

dew	humidity	temperature C 2m	calc_humidity
3.7	88.2	5.5	88.2
5.5	90.6	6.9	90.8
3.9	87.5	5.8	87.6
3.6	91	4.9	91.3
4.5	89.1	6.2	88.9
3.4	88.2	5.2	88.2

Вычисленное значение средней абсолютной ошибки составило 0.68, то есть погрешность вычисления будет <1%, что позволяет без ограничений использовать данный метод восстановления значений относительной влажности.

Стоит отметить, что восстанавливать будем те значения, где присутствуют данные о температуре точки росы (то есть те, где температура точки росы не равна 0).

Рассматривая новый график плотности распределения значений относительной влажности (рис. 3.6), можно говорить о небольшом увеличении качества исходных данных.

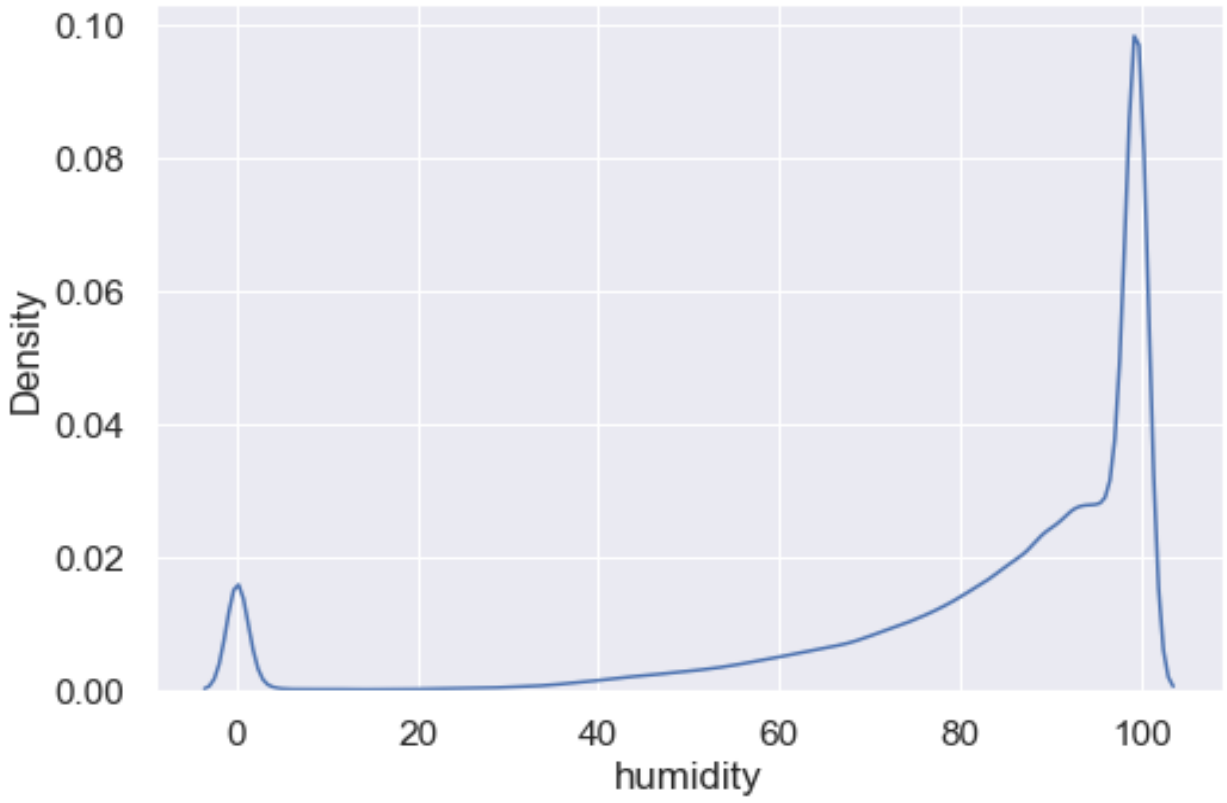


Рисунок 3.6 – График плотности распределения значений относительной влажности после восстановления данных

3.2.4 Температура на разных уровнях

Данные о температуре на разных уровнях будем рассматривать комплексно. Для начала проанализируем данные о температуре на высоте 0 см (рис. 3.7)

Температура на высоте 0 см (фактически температура поверхности дорожного полотна) не имеет явных выбросов или неверных значений.Пик на нулевом значении скорее всего говорит нам о том, что неизвестные значения температуры так же заменялись нулевыми значениями, но в данной ситуации это не оказывает существенного влияния.

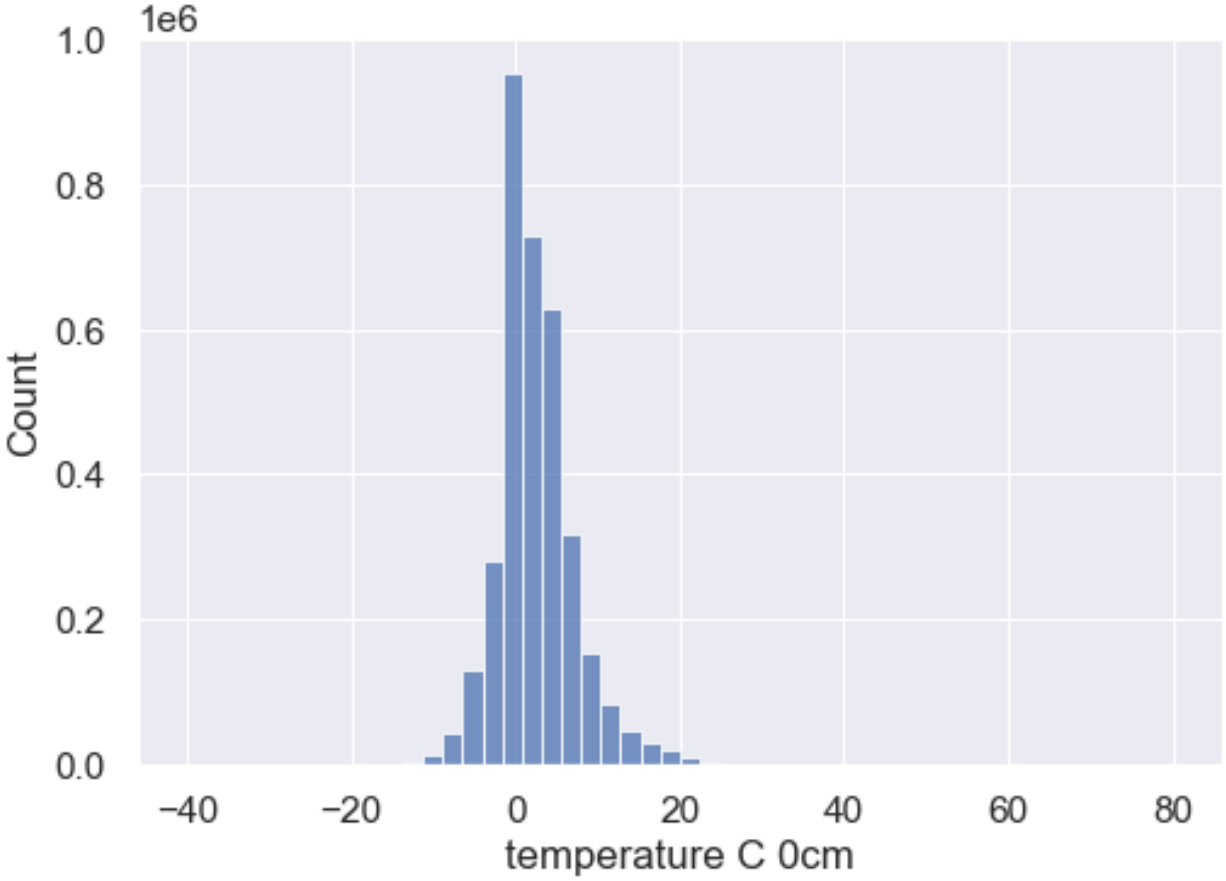


Рисунок 3.7 – Гистограмма распределения значений температуры на высоте 0cm

Также рассмотрим крайние значения:

- min = - 40 °C
- max = 80 °C

Очевидно, что подобные значения являются крайне нереалистичными, поэтому ограничим выборку по значению 0.05 и 0.95 квантилей соответственно.

Рассмотрим гистограмму распределения значений температуры на высоте 20 см (рис. 3.8)

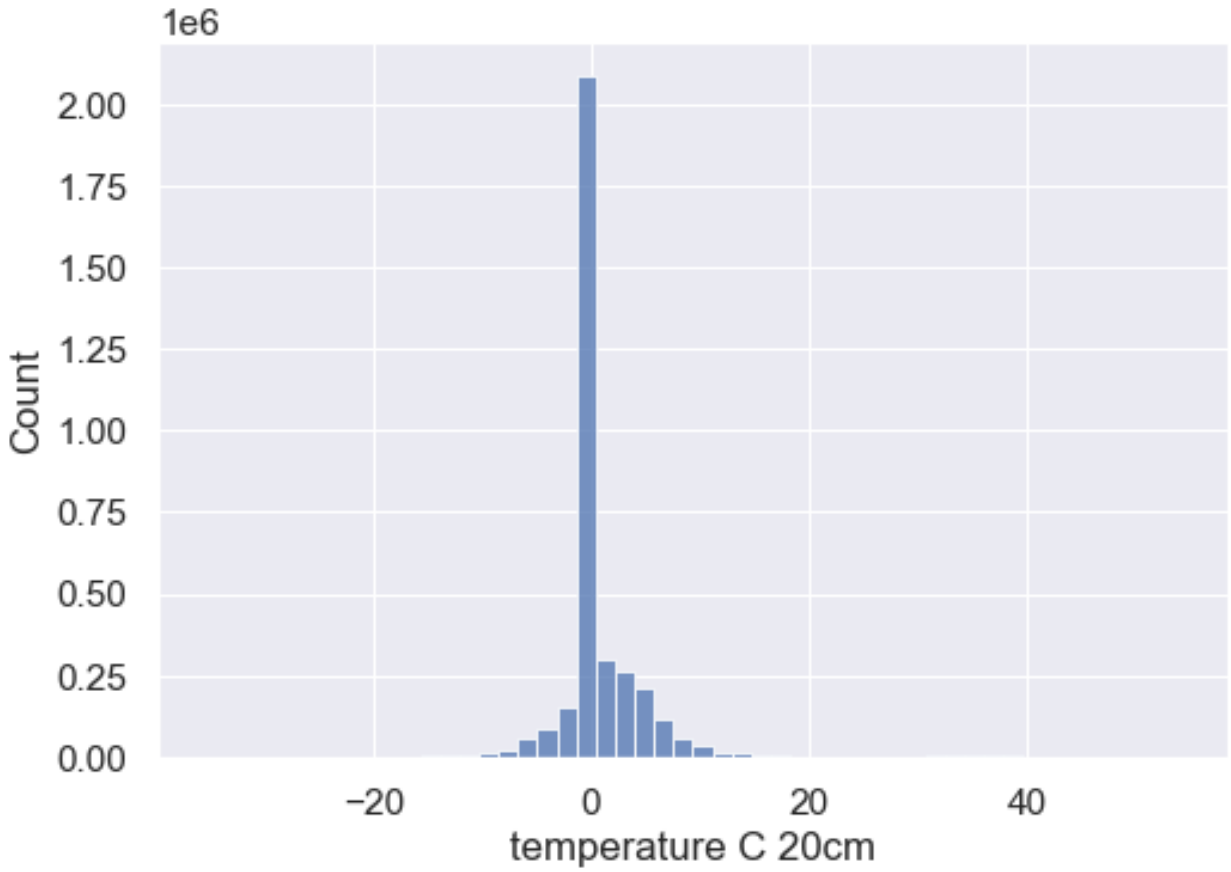


Рисунок 3.8 – Гистограмма распределения значений температуры на высоте 20cm

Температура на высоте 20 см имеет ярко выраженный пик нулевых значений.

Крайние значения:

- min = - 35.2 °C
- max = 53.8 °C

Восстановление данных значений нецелесообразно, поэтому данный тип данных можно либо исключить из нашего набора, либо задать ему пониженные веса.

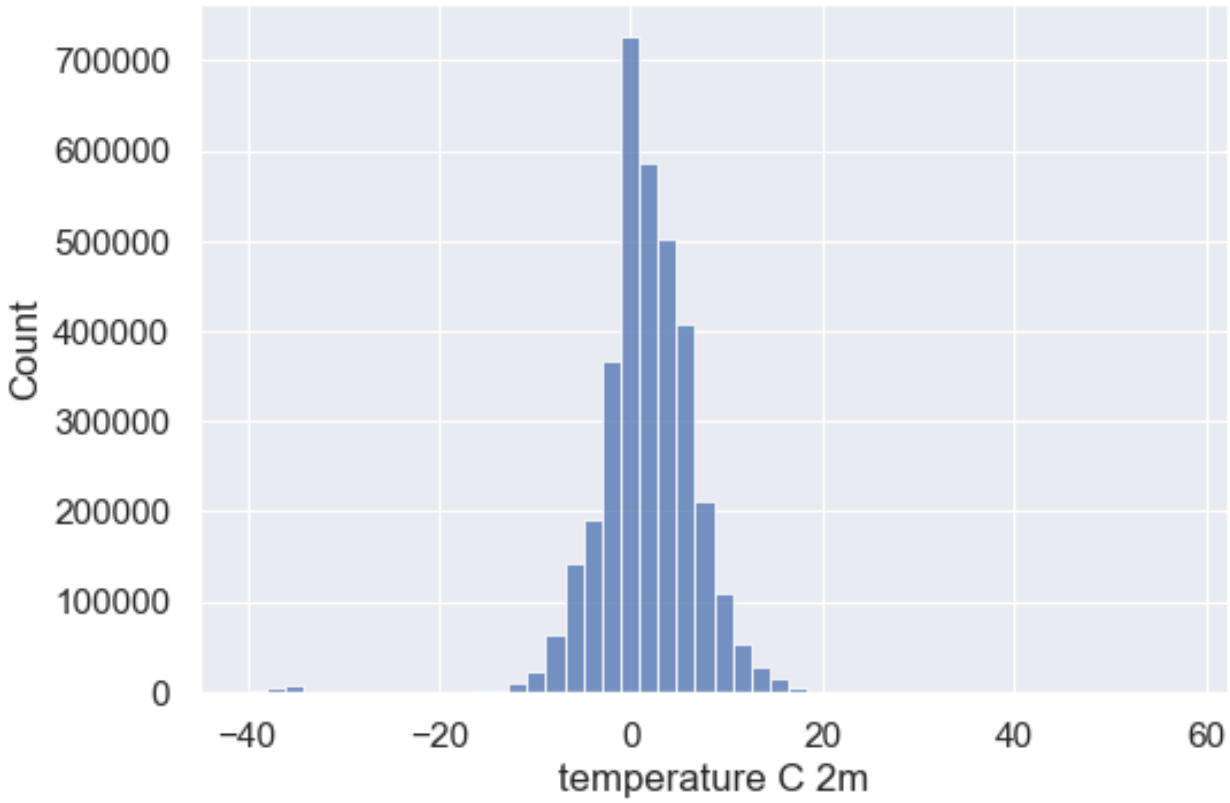


Рисунок 3.9 – Гистограмма распределения значений температуры на высоте 2m

Температура на высоте 2 м (рис. 3.9.) не имеет явных выбросов или неверных значений. Восстановленные нулевые данные так же не оказывают сильного влияния, крайние значения:

- min = - 40.0 °C
- max = 57.3 °C

Мы так же ограничим их по значениям 0.05 и 0.95 квантилей.

Итого для обучения модели мы будем использовать данные о температуре на высотах 0 см и 2 м.

3.2.3.Направление и скорость ветра

Рассматривая гистограммы распределения значений скорости и направления ветра (рис. 3.10, 3.11), можно сделать следующие выводы:

- Пропущенные значения так же заменялись нулевыми значениями

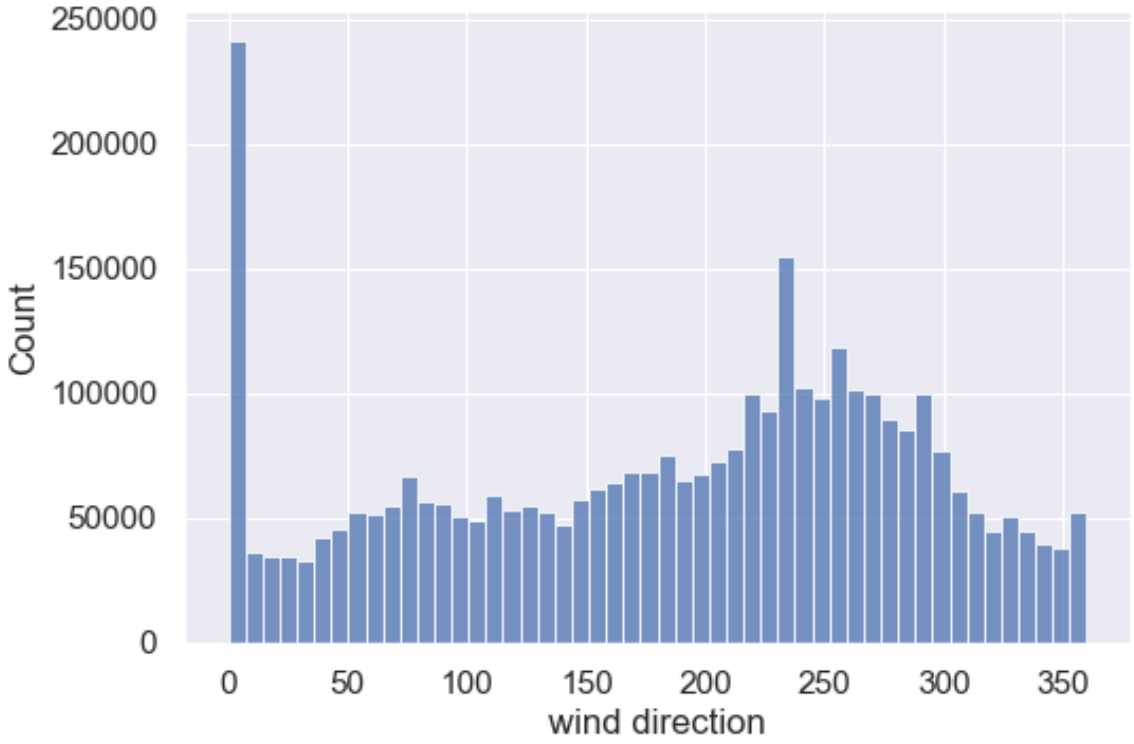


Рисунок 3.10 – Гистограмма распределения значений направления ветра

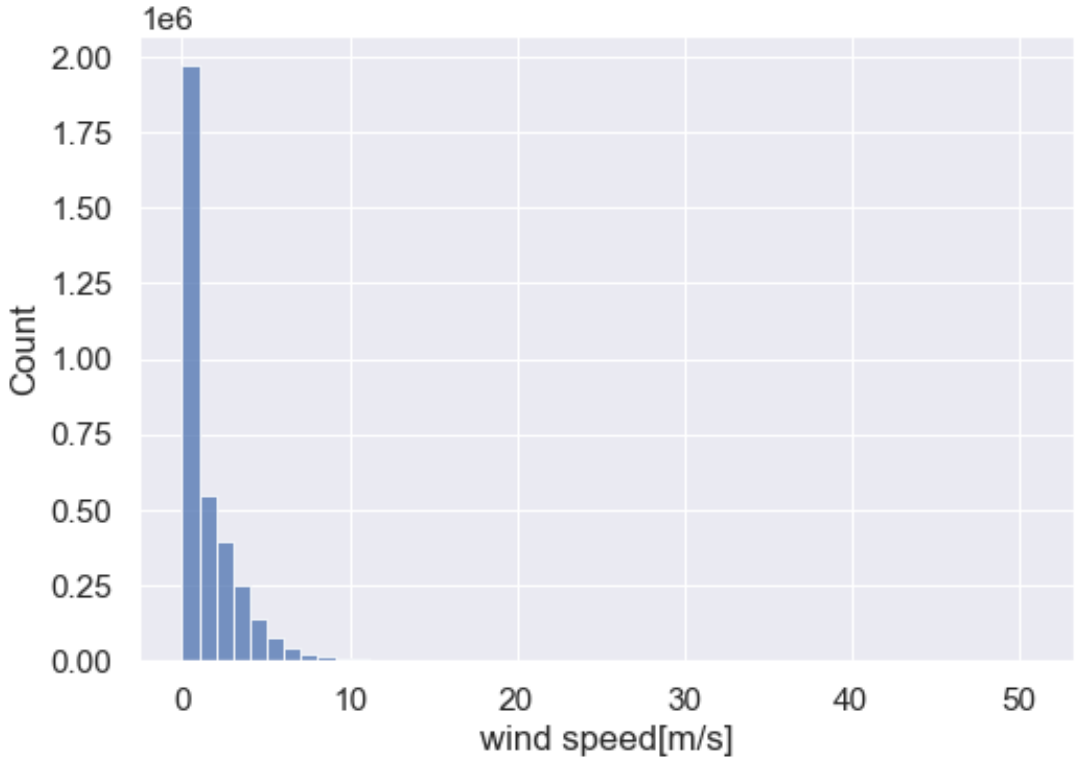


Рисунок 3.11 – Гистограмма распределения значений скорости ветра

- Если исключить из расчета нулевые значения, то преобладающим направлением ветра станет юго-западное.
- Если исключить из расчета нулевые значения (рис. 3.12), то преобладающими скоростями ветра все равно останутся значения <1 м/с.

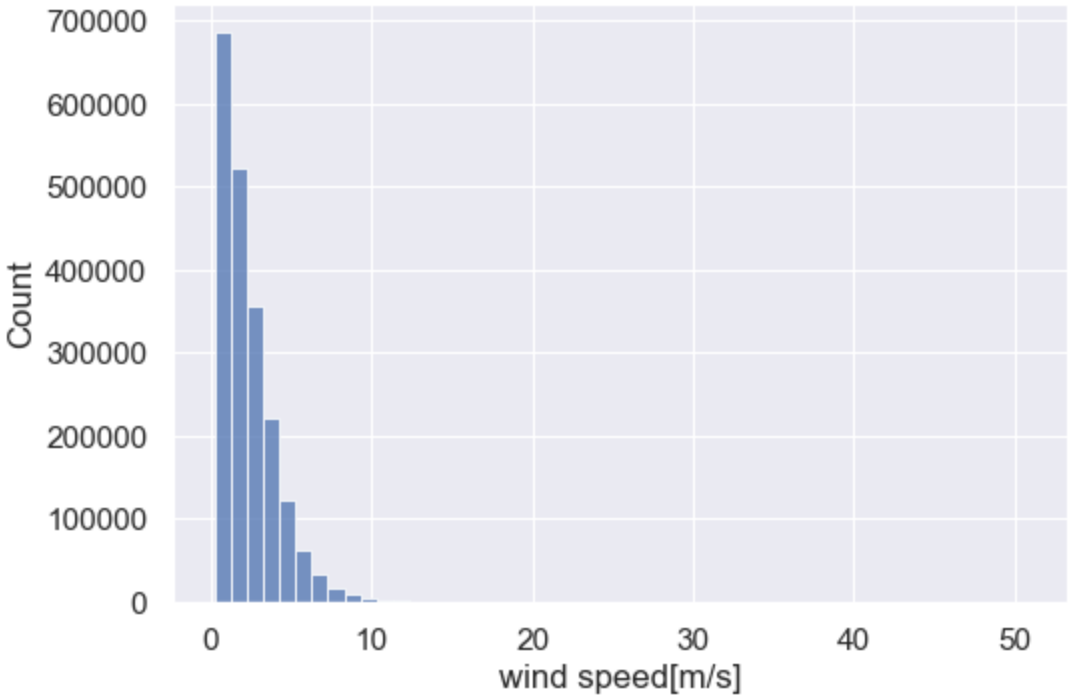


Рисунок 3.12 – Гистограмма распределения значений скорости ветра без 0 значений

Не будем применять никаких ограничений на использование этих данных.

3.2.4 Осадки

Теперь рассмотрим данные о количестве осадков (рис. 3.13).

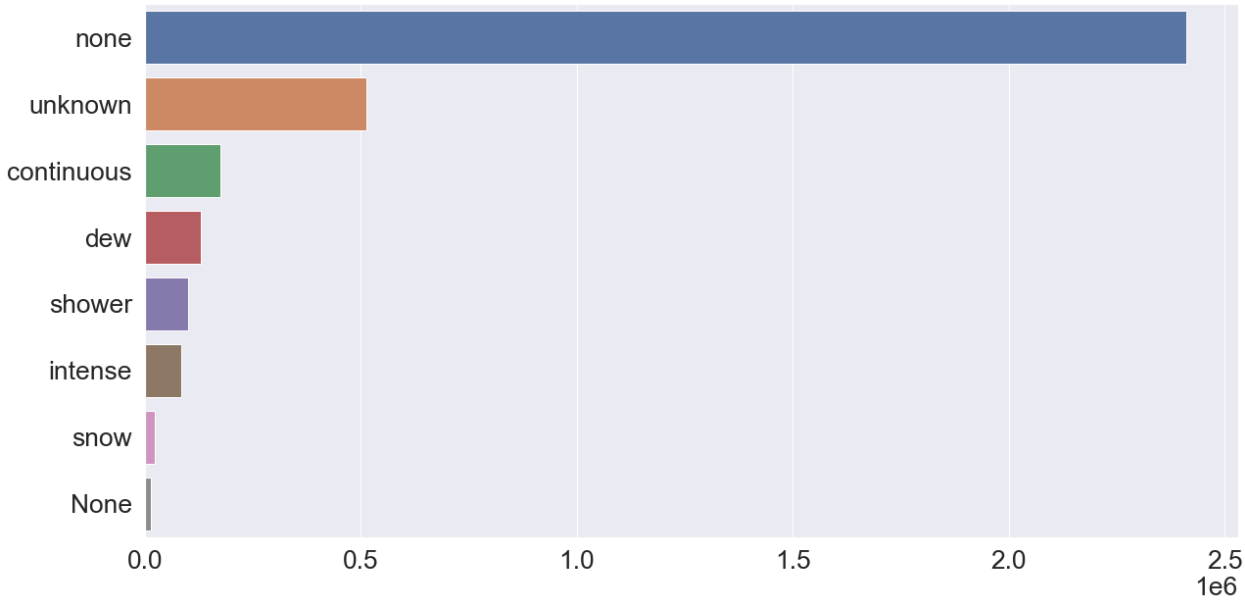


Рисунок 3.13 – Диаграмма распределения типов осадков

В наборе данных имеется очень большое количество значений “none”. Но в данной выборке присутствует класс ‘unknown’, к которому присваивались неизвестные значения. Поэтому значения ‘none’ являются действительными. Так же имеется класс ‘None’, идентичный первому. Объединим их.

Рассмотрим количество экземпляров:

- none - 2410366
- unknown - 512883
- continuous - 174924
- dew - 130318
- shower - 99382
- intense - 84601
- snow - 22969
- None – 14236

Мы также можем наблюдать сильный перекос в сторону класса ‘none’, но так как значение типов осадков имеют наибольшую связь со значениями

нашей целевой переменной, а именно roadCondition, то мы ограничимся балансировкой именно целевой переменной.

Результаты “очистки” нашего набора данных и последующего восстановления баланса с помощью oversampling’а можно наблюдать на (рис. 3.14). Значений “опасных” классов все еще меньше, чем “безопасных”, но перекос в их сторону стал существенно меньше, что позволит в последующем увеличить качество нашей модели:

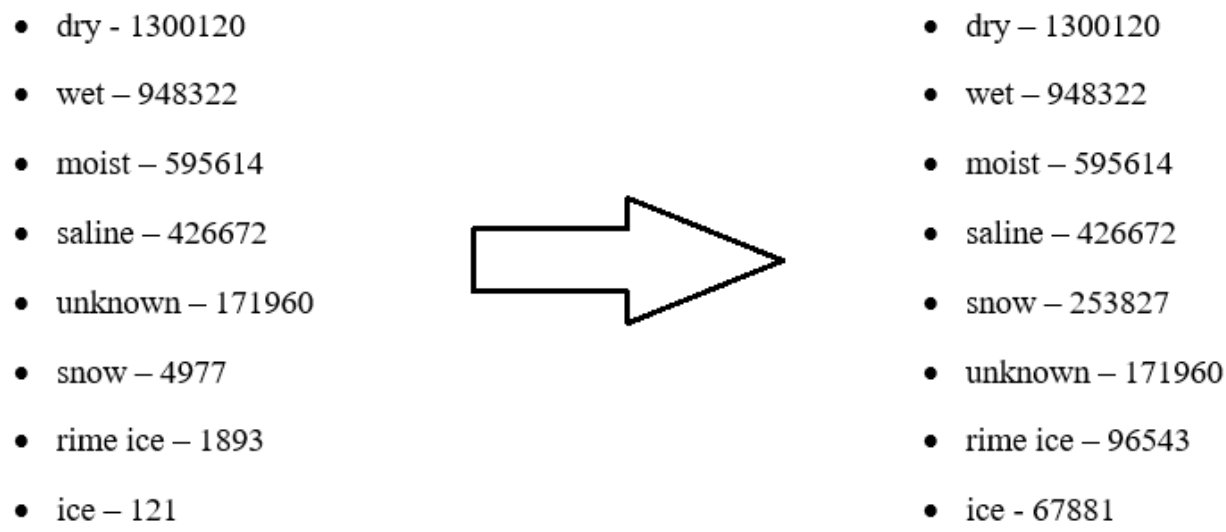


Рисунок 3.14 – Результаты балансировки

3.2.5 Восстановление данных

В нашем наборе данных имеется большое количество экземпляров класса ‘unknown’ –более 170 тыс. значений с неизвестным состоянием поверхности. Мы можем попытаться восстановить их, используя имеющиеся данные об осадках, температуре, влажности и ветре.

Будем решать задачу мультиклассовой классификации:

- ✓ Удалим из набора данные с неизвестными значениями целевой переменной для последующего восстановления.
- ✓ Определим массивы для обучения:

- X – dew, humidity, precipitation, temperature C 0cm, temperature C 2m, wind direction, wind speed[m/s]
- y – roadCondition

✓ Для обучения закодируем значения категориальных переменных (рис. 3.15) (precipitation)с помощью быстрого кодирования.

precipitation
none
continuous
dew
shower
intense
snow

continuous	dew	intense	none	shower	snow
0	0	0	1	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	0	0	0	1	0
0	0	1	0	0	0
0	0	0	0	0	1

Рисунок. 3.15 – Принцип быстрого кодирования

Разобьем наш набор данных на обучающую и тестовые выборки в соотношении 0.33 / 0.67 (тестовая и обучающая соответственно).

Для обучения использовались CatBoostClassifier, LGBMClassifierиXGBClassifier со стандартными входными параметрами, ограниченные только по количеству итераций (300, 500, 700, 1200).

Наилучший результат показал LGBMClassifier с количеством итераций 700:

AC cat = 0.64

AC xgb = 0.635

AC lgb = 0.689

AC (Accuracy) – метрика точности, отношение количества правильных предсказаний к общему числу предсказаний.

Применяем модель для восстановления значений состояния покрытия. Распределение полученных значений (рис 3.16.):

- wet - 128445

- saline - 23081
- dry - 16630
- moist - 2900
- snow - 904

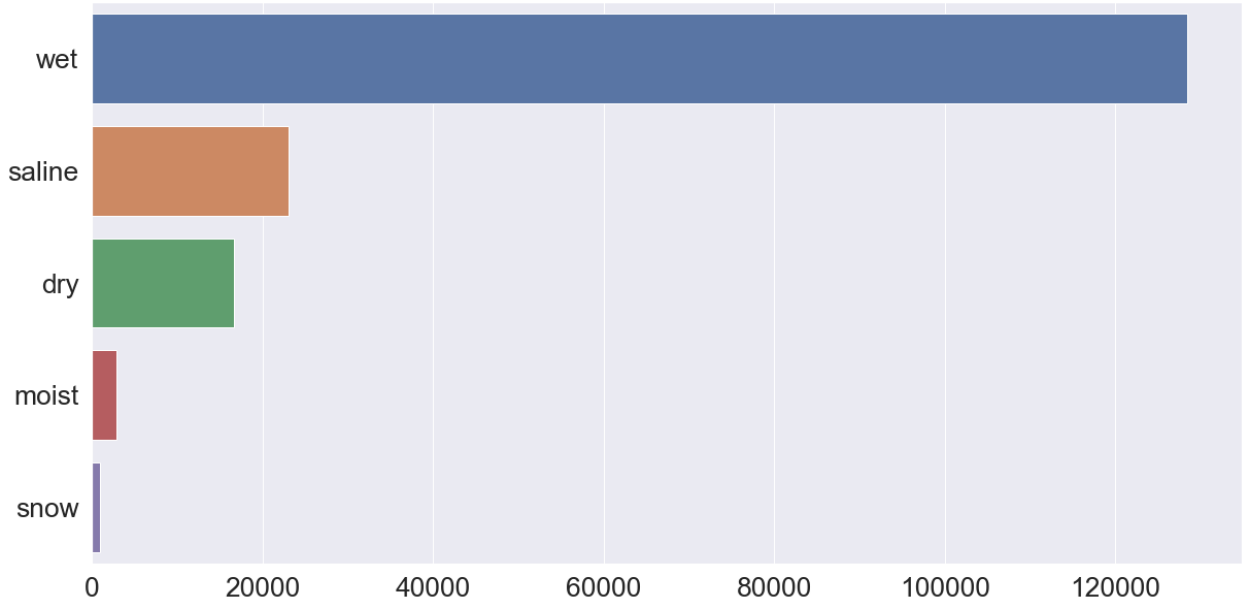


Рисунок 3.16 – Диаграмма распределения предсказанных классов

4 Прогноз состояния дорожного покрытия

Для прогноза значения состояния дорожного покрытия будем использовать несколько разных конфигураций. В основе прогноза будем использовать временной лаг. Таблицы формируются по следующему принципу (рис. 3.17.):

num	datetime	X_value	y_value
n1	2019-03-12 23:00:00	x1	y1
n2	2019-03-12 23:20:00	x2	y2
n3	2019-03-12 23:40:00	x3	y3
n4	2019-03-13 00:00:00	x4	y4

num	datetime	X_value_1	X_value_2	X_value_3	y_value
n4	2019-03-12 23:40:00	x1	x2	x3	y4

Рисунок 4.1 – Принцип формирования обучающей выборки

1. Выборка разделяется по станциям, чтобы иметь не смешиваемые временные ряды.
2. Выбирается временной промежуток, который будет доступен модели для предсказания. Для нас это 100 минут и 5 часов.
3. Из 1 колонки параметров формируется 5 и 15 колонок для каждого периода соответственно
4. Целевое значение смещается на определенное количество записей. Для нас это 20 минут, 1 час и 3 часа.
5. В итоге мы получаем обучающую выборку, где в каждой строке представлены данные за предыдущий период и значение целевой переменной через 20 минут, 1 час и 3 часа соответственно.

Выборки были разделены на обучающую и тестовую в соотношении 0.33/0.67

Для решения задачи был выбран CatBoostClassifierco стандартными параметрами и количеством итераций равным 300

После обучения модели, предсказания значений тестовой выборки и последующего сравнения предсказанных значений были получены следующие значения метрики Accuracy (точность) для соответствующих конфигураций:

Полученные результаты:

- для базового временного ряда в 100 минут и прогноза состояния покрытия через 20 минут: AC = 0.89
- для базового временного ряда в 100 минут и прогноза состояния покрытия через 1 час:AC = 0.84
- для базового временного ряда в 100 минут и прогноза состояния покрытия через 3 часа:AC = 0.76

Полученные значения можно интерпретировать буквально так:

Используя временной ряд в 100 минут и прогнозируя значение состояния поверхности дорожного полотна через 20 минут, прогноз будет верен в 89% случаев.

Полученные значения можно интерпретировать буквально так:

Используя временной ряд в 100 минут и прогнозируя значение состояния поверхности дорожного полотна через 1 час, прогноз будет верен в 84% случаев.

Полученные значения можно интерпретировать буквально так:

Используя временной ряд в 100 минут и прогнозируя значение состояния поверхности дорожного полотна через 3 часа, прогноз будет верен в 76% случаев.

Заключение

Исходя из результатов проведенного исследования, а именно из довольно высоких показателей точности тестовых моделей, а также высокого потенциала усовершенствования моделей мы можем сделать некоторые выводы.

Во-первых, прогнозирование состояния дорожной поверхности поможет лицам, ответственным за состояние дорожной обстановки принимать решения с большей уверенностью и точностью, например об использовании специальных средств для предотвращения наступления опасного состояния дорожного покрытия, об оперативном улучшении состояния дорожного покрытия (уборка и расчистка снега с помощью специализированной техники и устройств), или же о простом предупреждении участников дорожного движения о скором ухудшении или улучшении ситуации на дороге.

Во-вторых, исходя из предыдущего пункта, помощь в принятии решений позволит сократить расходы на содержание и обслуживание дорожной инфраструктуры, например, уменьшая количество ложно положительных выездов уборочной техники, когда потраченные средства не приносят ожидаемого эффекта. Также это поможет снизить операционные расходы на прогнозы, составляемые человеком.

В-третьих, использование моделей машинного обучения в подобной сфере деятельности является очень перспективным, благодаря возможности подстраиваться под очень большое количество факторов, в том числе индивидуальных для каждого единичного экземпляра (участка дороги), такие как, например, высота поверхности над уровнем моря, процентное соотношение типов растительности вокруг исследуемого участка, наличие и расстояние до ближайших водоемов, низин и возвышенностей, величина и тип трафика в разные времена года, месяца, недели, дня или даже часа, а

также сотни других признаков, которые можно извлечь из единичного экземпляра.

В-четвертых, одним из самых главных выводов является то, что благодаря превентивным мерам, будь то уборка снега, обработка дорожного полотна, или даже простое предупреждение, сможет снизиться количество жертв дорожно-транспортных происшествий, раненых и погибших, что несомненно очень важно для любого гражданского общества.

Для дальнейшего развития этого направления, и создания моделей для промышленного использования (использования в реальных условиях, на реальных дорогах) можно предложить следующие пути и элементы усовершенствования моделей:

- Использовать архивные данные за больший период времени. Данные в нашем исследовании были представлены довольно короткой выборкой длиной в 5 месяцев. Обучая модель на большем количестве записей, мы сможем находить скрытые закономерности, а также годовые, месячные, дневные ходы целевой переменной
- Использовать индивидуальные данные для конкретной географической точки. На значение целевой переменной могут серьезно влиять географические показатели, причем как базовые (например, широта, или климатический пояс), так и точечные (преобладающий тип растительности, или расстояние до водоемов), различные климатические показатели и т.д. Поэтому для обучения важно использовать максимальное количество доступных признаков, для выявления всех скрытых закономерностей.
- Использовать данные соседних станций, а также прочих станций, похожих по тем или иным признакам. Это может быть не только схожий тип климата, но и интенсивность траффика, тип и

толщина покрытия, причем другие признаки могут существенно отличаться.

- Использовать комбинации признаков и последующий их отбор. Для промышленной модели важна не только точность, но и скорость прогноза. Использование комбинаций признаков, причем как простых (например, произведение), так и сложных (сложная функция от комбинации определенных признаков с условиями), позволяет увеличить значимость признаков, производя их последующий отбор и увеличивая скорость работы и простоту модели.
- Использовать актуальные синоптические данные о положении барических образований. Это позволит обосновать прогноз не только временными рядами различных показателей, но и, например, данными гидродинамических прогнозов погоды для данной местности.
- Использовать ансамбли моделей для получения наилучшего результата. Суть заключается в том, чтобы не использовать все найденные признаки в одной полученной модели, а правильно распределить их по нескольким, тем самым повысив не только скорость, но и эффективность и точность. Разные модели могут использовать данные друг друга, иметь различную частоту отработки и т.д.

В заключение хочется отметить, что изучение и использование машинного обучения для решения различных метеорологических задач является очень важным условием развития данной области в современное время, и этому должно уделяться куда большее внимание, чем уделяется сейчас.

Список использованных источников

1. Отраслевой дорожный методический документ методические рекомендации по специализированному прогнозу состояния дорожного покрытия, ОДМ 218.2.003-2009
2. Отраслевой дорожный методческий документ руководство по борьбе с зимней скользкостью на автомобильных дорогах, Министерство Транспорта Российской Федерации, Государственная служба дорожного хозяйства (РОСАВТОДОР), Москва 2003
3. Отраслевой дорожный методический документ методические рекомендации по специализированному гидрометеорологическому обеспечению дорожного хозяйства, ОДМ 218.8.001-2009
4. Best Practices for Road Weather Management, USDOT FHWA May 2003
5. Roadway Icing and Weather: A Tutorial, Department of Atmospheric Sciences, University of Washington 2020
6. Учебник по машинному обучению, ШАД 2022
7. Машинное обучение с использованием Python, Крис Элбон
8. https://atmos.uw.edu/ - Сайт департамента атмосферных наук Вашингтонского университета. Дата обращения – 03.05.2022.
9. https://ops.fhwa.dot.gov/ - Сайт Министерства транспорта США. Дата обращения – 03.05.2022.
10. https://dataverse.harvard.edu/ - Сайт банка данных Гарвардского университета. Дата обращения – 03.05.2022.
11. https://forecast.weather.gov/ - Сайт национальной метеорологической службы США. Дата обращения – 03.05.2022.
12. http://stat.gibdd.ru/ - Сайт со статистическими данными ГИБДД. Дата обращения – 03.05.2022.
13. https://ml-handbook.ru/ - Справочник по машинному обучению. Дата обращения – 03.05.2022.

14. https://habr.com/ - Сайт с публикациями аналитических статей, связанных с информационными технологиями. Дата обращения – 03.05.2022.
15. http://www.machinelearning.ru/ - Профессиональный информационно-аналитический ресурс, посвященный машинному обучению. Дата обращения – 03.05.2022.