**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ**

**Кафедра интеллектуальных систем**

РАКОВЕЦ

Андрей Владимирович

**АНАЛИЗ ЗНАЧИМОСТИ ВХОДНЫХ ПРИЗНАКОВ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ**

Дипломная работа

Научный руководитель:

старший преподаватель

Щетько Н.Н.

Допущен к защите

<<\_\_\_>> \_\_\_\_\_\_\_\_\_2022.

Зав. кафедрой интеллектуальных систем

кандидат физ.-мат. наук, доцент Козлова В.И.

Минск,2022.

**ОГЛАВЛЕНИЕ**

[Реферат дипломной работы 3](#_Toc104799935)

[ВВЕДЕНИЕ 6](#_Toc104799936)

[ГЛАВА 1 МОДЕЛИ С РЕШЕННОЙ ЗАДАЧЕЙ ОЦЕНКИ ЗНАЧИМОСТИ ПРИЗНАКОВ 7](#_Toc104799937)

[1.1 Бинарные деревья 7](#_Toc104799938)

[1.2 Случайные леса 15](#_Toc104799939)

[1.3 Перестановочная важность 18](#_Toc104799940)

[1.4 Суррогатные модели 20](#_Toc104799941)

[ГЛАВА 2 МЕТОДЫ ОЦЕНКИ ЗНАЧИМОСТИ ПРИЗНАКОВ В ЗАДАЧЕ ОТБОРА ПРИЗНАКОВ 23](#_Toc104799942)

[2.1 Задача отбора признаков 23](#_Toc104799943)

[2.2 Методы – фильтры(filters) 25](#_Toc104799944)

[2.3 Методы-обертки (Wrapper methods) 34](#_Toc104799945)

[2.4 Встроенные методы (embedded) 38](#_Toc104799946)

[ГЛАВА 3 ПРИМЕНЕНИЕ МЕТОДОВ ОЦЕНКИ ВАЖНОСТИ ПРИЗНАКОВ В РЕАЛЬНОЙ ЗАДАЧЕ 42](#_Toc104799947)

[3.1 Описание задачи 42](#_Toc104799948)

[3.2 Обучение моделей 42](#_Toc104799949)

[3.3 Внесение дополнительных шумовых признаков 48](#_Toc104799950)

[3.4 Подбор гиперпараметров 53](#_Toc104799951)

[ЗАКЛЮЧЕНИЕ 59](#_Toc104799952)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 60](#_Toc104799953)

# Реферат дипломной работы

АНАЛИЗ ЗНАЧИМОСТИ ВХОДНЫХ ПРИЗНАКОВ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ

Дипломная работа объемом 63 страницы содержит 2 таблицы, 34 формулы, список использованной литературы из 38 наименований.

Целью исследования является исследование возможности интерпретации поведения обученных моделей машинного обучения с учителем на основе методов оценки важности входных признаков.

В качестве задач исследования, вытекающих из цели, ставилось рассмотреть проблему "черного ящика" при построении моделей машинного обучения с учителем, рассмотреть подход к анализу значимости признаков в моделях на основе деревьев принятия решений, рассмотреть задачу определения значимости признаков в контексте понижения размерности, исследовать алгоритмы на устойчивость к шумовым признакам с различными видами распределений.

Методы исследования: аналитический, статистический.

Новизна заключается в модификации одного из методов оценки важности признаков. Проведенные исследования показали, что данный модифицированный метод склонен увеличивать оценки важности признаков, которые были признаны маловажными оригинальным методом.

Ключевые слова: оценка важности входных признаков, отбор признаков, дерево решений, случайный лес, методы-фильтры, обёрточные методы, встроенные методы.

**Рэферат дыпломнай працы**

АНАЛІЗ ЗНАЧНАСЦІ ЎВАХОДНЫХ ПРЫКМЕТ МАДЭЛЯЎ МАШЫННАГА НАВУЧАННЯ З НАСТАЎНІКАМ

Дыпломная праца аб'ёмам 62 старонкі змяшчае 2 табліцы, 34 формулы, Спіс выкарыстанай літаратуры з 38 найменняў.

Мэтай даследавання з'яўляецца вывучэнне магчымасці інтэрпрэтацыі паводзін навучаных мадэляў машыннага навучання з настаўнікам на аснове метадаў ацэнкі важнасці ўваходных прыкмет.

У якасці задач даследавання, якія вынікаюць з мэты, ставілася разгледзець праблему "чорнай скрыні" пры пабудове мадэляў машыннага навучання з настаўнікам, разгледзець падыход да аналізу значнасці прыкмет у мадэлях на аснове дрэў прыняцця рашэнняў, разгледзець задачу вызначэння значнасці прыкмет у кантэксце паніжэння памернасці, даследаваць алгарытмы на ўстойлівасць да шумавых прыкметах з рознымі відамі размеркаванняў.

Метады даследавання: аналітычны, статыстычны.

Навізна заключаецца ў мадыфікацыі аднаго з метадаў ацэнкі важнасці прыкмет. Праведзеныя даследаванні паказалі, што дадзены мадыфікаваны метад схільны павялічваць ацэнкі важнасці прыкмет, якія былі прызнаныя малаважнымі арыгінальным метадам.

Ключавыя словы: ацэнка важнасці ўваходных прыкмет, адбор прыкмет, дрэва рашэнняў, выпадковы лес, метады-фільтры, абгортачныя метады, убудаваныя метады.

Summary of the thesis

ANALYSIS OF THE SIGNIFICANCE OF INPUT FEATURES OF MACHINE LEARNING MODELS WITH A TEACHER

The 63-page thesis contains 2 tables, 34 formulas, a list of references from 38 titles.

The aim of the study is to investigate the possibility of interpreting the behavior of trained machine learning models with a teacher based on methods for assessing the importance of input features.

As research tasks arising from the goal, it was set to consider the problem of a "black box" when building machine learning models with a teacher, to consider an approach to analyzing the significance of features in models based on decision trees, to consider the task of determining the significance of features in the context of dimension reduction, to investigate algorithms for resistance to noise features with different types of distributions.

Research methods: analytical, statistical.

The novelty lies in the modification of one of the methods for assessing the importance of features. The conducted studies have shown that this modified method tends to increase the estimates of the importance of features that were considered unimportant by the original method.

Keywords: evaluation of the importance of input features, feature selection, decision tree, random forest, filter methods, wrapper methods, embedded methods.

# ВВЕДЕНИЕ

Одна из фундаментальных проблем машинного обучения в том, что обученные модели работают как "черный ящик". Машинное обучение в целом сводится к тому, что какая-то очень гибкая и настраиваемая функцию подгоняем под такой вид, который хорошо описывает какие-то имеющиеся данные. Например, зная, как формируется нейронная сеть, зная, как происходит её обучение, но после того, как она обучена, её поведение очень сложно интерпретировать - она может имеющиеся у нас данные описывать просто идеально, но информация о том, какие именно принципы лежат в полученном функциональном преобразовании, будет по-прежнему неизвестна. Т.е. зная выходной результат, исследователь не знает, почему получил именно такой результат, а не какой-то другой.

Цель: исследовать возможность интерпретации поведения обученных моделей машинного обучения с учителем на основе методов оценки важности входных признаков для модели.

Задачи:

* рассмотреть проблему "черного ящика" при построении моделей машинного обучения с учителем,
* рассмотреть подход к анализу значимости признаков в моделях на основе деревьев принятия решений,
* рассмотреть задачу определения значимости признаков в контексте понижения размерности,
* исследовать алгоритмы на устойчивость к шумовым признакам с различными видами распределений

# ГЛАВА 1 МОДЕЛИ С РЕШЕННОЙ ЗАДАЧЕЙ ОЦЕНКИ ЗНАЧИМОСТИ ПРИЗНАКОВ

## 1.1 Бинарные деревья

Рассмотрим бинарное дерево, в котором:

* каждой внутренней вершине *v* приписана функция (предикат)   
  *βv* : X →{0*,* 1};
* каждой листовой вершине v приписан прогноз (в случае с классификацией листу также может быть приписан вектор вероятностей).

Рассмотрим алгоритм р(x), который начинает свое выполнение из корневой вершины v0 и вычисляет значение функции βv0. Если оно равно нулю, то алгоритм переходит в левую дочернюю вершину, если единице - в правую, вычисляет значение предиката в новой вершине и повторяет данную процедуру. Процесс продолжается, пока не будет достигнута листовая вершина. Данный алгоритм возвращает класс, приписанный данной вершине. Такой алгоритм называется бинарным решающим деревом и описан в работе [39].

На практике в большинстве случаев используются одномерные предикаты βv, которые сравнивают значение одного из признаков с числом:

*βv*(*x*; *j, t*) = [*xj < t*]*.*

Существуют и многомерные предикаты, например:

* + линейные *βv*(*x*) = [⟨*w, x*⟩ *< t*];
  + метрические *βv*(*x*) = [*ρ*(*x, xv*) *< t*], где точка *xv* является одним из объектов выборки - некоторой точкой признакового пространства.

Многомерные предикаты позволяют получить более сложные разделяющие гиперповерхности, но редко используются на практике — например, из-за того, что увеличивают и без того огромные риски переобучения данных деревьев. Далее будем рассматривать только одномерные предикаты.

Очевидно, что для любой выборки существует такое решающее дерево, которое не допускает на ней ни одной ошибки — дерево, в каждом листе которого находится ровно по одному объекту выборки. Высока вероятность того, что это дерево окажется переобученным и не сможет показать хорошую точность предсказаний на новых данных. Можно было бы поставить задачу поиска такого дерева, которое является минимальным с точки зрения количества листовых вершин среди всех решающих деревьев, которые допускают минимум ошибок на обучении — в таком случае можно было бы рассчитывать на наличие у дерева некоторой обобщающей способности. К сожалению, данная задача является NP-полной, и поэтому приходится ограничиваться жадными алгоритмами построения дерева [7].

Опишем базовый алгоритм построения бинарного решающего дерева. Начиная со всей обучающей выборки *X* найдем наилучшее ее разбиение на две части *R*1(*j, t*) = {*x* | *xj < t*} и *R*2(*j, t*) = {*x* | *xj* ≥ *t*} с точки зрения заранее заданного функционала качества *Q*(*X, j, t*) - сравнительного качества различных способов разбиения заданной совокупности элементов на классы, т. е. того количественного критерия, следуя которому можно было бы предпочесть одно разбиение другому. Найдя наилучшие значения *j* и *t*, на выделенной основе создадим корневую вершину дерева, поставив ей в соответствие предикат [*xj <t*]. Объекты разобьются на две (для бинарного дерева) группы — одни для левого поддерева, другие для правого. Для каждой из указанных выше групп рекурсивно повторим процедуру, построив дочерние вершины для корневой, и так далее. В каждой вершине проверим выполнение некоторого условия останова — при выполнении останавливаем рекурсию и объявляем данную вершину листом. Когда дерево построено, каждому листу ставится в соответствие некоторая выходная метрика. В случае с классификацией это может быть класс, к которому относится больше всего объектов в листе, или вектор вероятностей (например, вероятность класса может быть равна доле его объектов в листе). Для регрессии это может быть среднее значение, медиана или некоторая другая функция от целевых переменных объектов в листе. Выбор конкретной функции зависит от функционала качества в исходной задаче [39].

Решающие деревья могут обрабатывать пропущенные значения — ситуации, в которых для некоторых объектов неизвестны значения одного или нескольких признаков. Для этого необходимо модифицировать процедуру разбиения выборки в вершине, что можно сделать несколькими способами [7].

После того, как дерево построено, можно провести стрижку (pruning) — удаление некоторых вершин с целью понижения сложности и повышения способности к обобщению. Существует несколько подходов к стрижке, о которых немного упомянем ниже.

Таким образом, конкретный метод построения решающего дерева определяется [7]:

1. Видом предикатов в вершинах;
2. Функционалом качества
3. Критерием останова;
4. Методом обработки пропущенных значений;
5. Методом стрижки.

Также могут иметь место различные расширения, связанные с учетом весов объектов, работой с категориальными признакам и т.д.

При построении дерева необходимо задать функционал качества, на основе которого осуществляется разбиение выборки на каждом шаге. Обозначим через *Rm* множество объектов, попавших в вершину, разбиваемую на данном шаге, а через *Rℓ* и *Rr* — объекты, попадающие в левое и правое поддерево соответственно при заданном предикате. Обычно используются использовать функционалы следующего вида [39]:

Здесь — это критерий информативности (impurity criterion), который оценивает качество распределения целевой переменной среди объектов множества . Чем меньше разнообразие целевой переменной, тем меньше должно быть значение критерия информативности — и, соответственно будем пытаться минимизировать его значение. Функционал качества при этом будем максимизировать [7]. Как уже обсуждалось выше, в каждом листе дерево будет выдавать метрику — вещественное число, вероятность или класс. Исходя из этого, можно оценивать информативность множества объектов тем, насколько хорошо их целевые переменные предсказываются константой (при оптимальном выборе данной константы):

где L (y, c) — некоторая функция потерь. Далее рассмотрим, какие именно критерии информативности часто используются в задачах регрессии и классификации.

Одним из распространённых критериев в задачах **регрессии** является квадрат отклонения в качестве функции потерь. В этом случае критерий информативности будет выглядеть как

Как несложно показать, минимум в этом выражении будет достигаться на среднем значении целевой переменной. Значит, критерий можно переписать в следующем виде:

Таким образом, информативность вершины измеряется её дисперсией — чем меньше разброс целевой переменной, тем лучше вершина. Разумеется, можно использовать и другие функции ошибки L — например, при выборе абсолютного отклонения в качестве критерия получается среднее абсолютное отклонение от медианы.

Для задач **классификации** будем использовать следующие обозначения [7]:

Обозначим через pk долю объектов класса k (k ∈ {1, . . ., K}), попавших в вершину R:

Через обозначим класс, чьих представителей оказалось больше всего среди объектов, попавших в данную вершину:

Рассмотрим индикатор ошибки как функцию потерь:

Легко видеть, что оптимальным предсказанием тут будет наиболее популярный класс — значит, критерий ошибки классификации будет равен следующей доле ошибок:

Данный критерий ошибки классификации является достаточно грубым, поскольку учитывает частоту pk∗ лишь одного класса.

Рассмотрим ситуацию, в которой выдаем вершине не один класс, а распределение на всех классах Качество такого распределения можно измерять, например, с помощью **критерия Бриера**:

Можно показать, что оптимальный вектор вероятностей состоит из долей классов pk [7]:

Если подставить данные вероятности в исходный критерий информативности и провести ряд преобразований, то получим **критерий Джини** [7]:

**Энтропийный критерий** является более популярным способом оценивания качества вероятностей — логарифмическими потерями, или логарифмом правдоподобия:

Для вывода оптимальных значений ck вспомним, что каждая вершина должна назначить объекту, попавшему в вершину, класс, поэтому все значения ck должны суммироваться в единицу. Как известного из методов оптимизации, для учёта этого ограничения необходимо искать минимум лагранжиана:

Дифференцируя, получаем:

откуда выражаем ck = pk/λ. Суммируя указанные выше равенства по k, получим

откуда λ = 1. Значит, минимум достигается при ck = pk, как и в предыдущем случае. Подставляя данные выражения в критерий, получим, что он будет представлять собой энтропию распределения классов [7]:

Из теории вероятностей известно, что энтропия ограничена снизу нулем, причем минимум достигается на вырожденных распределениях (pi = 1, pj = 0 для i ≠ j). Максимальное же значение энтропия принимает для равномерного распределения. Отсюда заметно, что энтропийный критерий будет отдавать предпочтение более вырожденным распределениям классов в вершине [7].

Можно придумать большое количестве **критериев останова**. Перечислим некоторые ограничения и критерии:

• Ограничение максимальной глубины дерева.

• Ограничение минимального числа объектов в листе.

• Ограничение максимального количества листьев в дереве.

• Останов в случае, если все объекты в листе относятся к одному классу.

• Требование, что функционал качества при дроблении улучшался как минимум на s процентов.

С помощью грамотного выбора подобных критериев и их параметров можно существенно повлиять на качество дерева. Тем не менее, такой подбор является трудозатратным и требует проведения большого количества оценок качества обучения при переборе..

**Стрижка дерева** является альтернативой критериям останова, описанным выше. При использовании стрижки сначала строится переобученное дерево (например, до тех пор, пока в каждом листе не окажется по одному объекту), а затем производится оптимизация его структуры с целью улучшения обобщающей способности. Существует ряд исследований, показывающих, что стрижка позволяет достичь лучшего качества по сравнению с ранним остановом построения дерева на основе различных критериев [7]. Тем не менее, на данный момент методы стрижки редко используются и не реализованы в большинстве библиотек для анализа данных. Причина заключается в том, что деревья сами по себе являются слабыми алгоритмами и не представляют большого интереса, а при использовании в композициях они либо должны быть переобучены (в случайных лесах), либо должны иметь очень небольшую глубину (в бустинге), из-за чего необходимость в стрижке отпадает.

Одним из методов стрижки является cost-complexity pruning. Обозначим дерево, полученное в результате работы жадного алгоритма, через *T0*. Поскольку в каждом из листьев находятся объекты только одного класса, значение функционала *R(T)* будет минимально на самом дереве *T0* (среди всех поддеревьев). Однако данный функционал характеризует лишь качество дерева на обучающей выборке, и чрезмерная подгонка под нее может привести к переобучению. Чтобы преодолеть эту проблему, введем новый функционал *Rα(T)*, представляющий собой сумму исходного функционала *R(T)* и штрафа за размер дерева:

где *|T|* — число листьев в поддереве *T*, а — параметр регуляризации. Это один из примеров регуляризованных критериев качества, которые ищут баланс между качеством классификации обучающей выборки и сложностью построенной модели. Можно показать, что существует последовательность вложенных деревьев с одинаковыми корнями:

(здесь *Tk* — тривиальное дерево, состоящее из корня дерева *T0*), в которой каждое дерево *Ti* минимизирует критерий (1) для α из интервала , причем

Эту последовательность можно достаточно эффективно найти путем обхода дерева. Далее из нее выбирается оптимальное дерево по отложенной выборке или с помощью кросс-валидации [7].

Одним из основных преимуществ решающих деревьев является возможность работы с **пропущенными значениями**. Рассмотрим некоторые варианты. Пусть нам нужно вычислить функционал качества для предиката β(x) = [xj < t], но в выборке R для некоторых объектов не известно значение признака j — обозначим их через *Vj* . В таком случае при вычислении функционала можно просто проигнорировать такие объекты, сделав поправку на потерю информации от этого:

Затем, если данный предикат окажется лучшим, поместим объекты из Vj как в левое, так и в правое поддерево. Также можно присвоить им при этом веса в левом поддереве и в правом. В дальнейшем веса можно учитывать, добавляя их как коэффициенты перед индикаторами [yi = k] во всех формулах. На этапе применения дерева необходимо выполнять похожий приём. Если объект попал в вершину, предикат которой не может быть вычислен из-за пропуска, то прогнозы для него вычисляются в обоих поддеревьях, и затем усредняются с весами, пропорциональными числу обучающих объектов в указанных поддеревьях. Иными словами, если прогноз вероятности для класса k в поддереве Rm обозначается через *amk(x)*, то получаем такую формулу:

Иной подход заключается в построении суррогатных предикатов в каждой вершине. Так называется предикат, который использует другой признак, но при этом дает разбиение, максимально близкое к данному. Отметим, что нередко схожее качество показывают и гораздо более простые способы обработки пропусков — например, можно заменить все пропуски на ноль. Для деревьев также разумно будет заменить пропуски в признаке на числа, которые превосходят любое значение данного признака. В этом случае в дереве можно будет выбрать такое разбиение по этому признаку, что все объекты с известными значениями пойдут в левое поддерево, а все объекты с пропусками — в правое [7].

Самый очевидный способ **обработки категориальных признаков** — разбивать вершину на столько поддеревьев, сколько имеется возможных значений у признака (multi-way splits) [7].

Рассмотрим подробнее другой подход. Пусть категориальный признак xj имеет множество значений *Q = {u1, . . ., uq}, |Q| = q*. Разобьем множество значений на два непересекающихся подмножества: *Q = Q1⊔Q2*, и определим предикат как индикатор попадания в первое подмножество: β(x) = [xj ∈ *Q1*]. Таким образом, объект будет попадать в левое поддерево, если признак xj попадает в множество Q1, и в правое поддерево в противном случае. Основная проблема заключается в том, что для построения оптимального предиката нужно перебрать 2q−1 − 1 вариантов разбиения, что может быть не вполне возможным.

С другой стороны, можно обойтись без полного перебора в случаях с бинарной классификацией и регрессией [1]. Обозначим через Rm() множество объектов, которые попали в вершину m и у которых j-й признак имеет значение , а через Nm() обозначим количество таких объектов. В случае с бинарной классификацией упорядочим все значения категориального признака на основе того, какая доля объектов с таким значением имеет класс +1:

после чего заменим категорию u(i) на число i, и будем искать разбиение как для вещественного признака. Можно показать, что если искать оптимальное разбиение по критерию Джини или энтропийному критерию, то получим такое же разбиение, как и при переборе по всем возможным 2q−1 − 1 вариантам [7,12].

## 1.2 Случайные леса

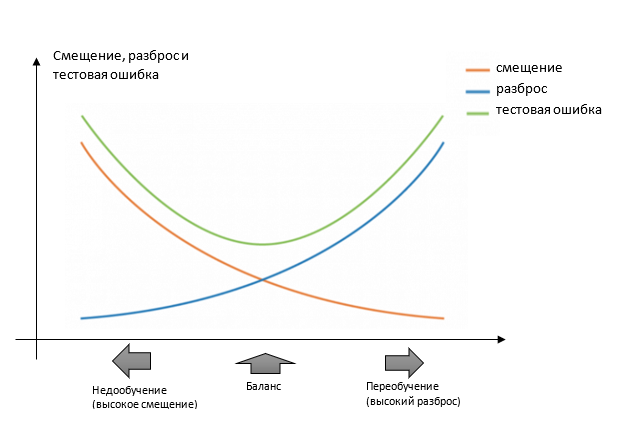
 Методы ансамбля — это парадигма машинного обучения, в которой несколько моделей (часто называемых "плохими учениками") обучаются для решения одной и той же проблемы и объединяются для достижения лучших результатов. Основная гипотеза состоит в том, что при правильном сочетании слабых моделей можно получить более точные и/или надежные модели [4]. Незначительные искажения и вариации модели, хотя она меняется в основном в противоположных направлениях, являются двумя наиболее основными характеристиками модели. Чтобы "решить" проблему, необходимо, чтобы наша модель имела достаточно степеней свободы, чтобы решить основную сложность используемых нами данных, но также необходимо, чтобы у нее было не слишком много степеней свободы, чтобы избежать его высокого разброса и быть более стабильной. Это известный компромисс между смещением и разбросом (рис. 1.1). [4].

Рисунок 1.1. - Иллюстрация компромисса между смещением и разбросом

В ансамблевой теории обучения вводятся понятия ***слабых учеников*** (или ***базовых моделей***), которые используются в качестве строительных блоков для проектирования более сложных моделей путем объединения некоторых из них. В большинстве случаев такие базовые модели работают сами по себе менее точно из-за того, что они имеют высокое смещение (например, модели с низкой степенью свободы), либо имеют слишком высокий разброс, чтобы быть устойчивыми (например, модели с высокой степенью свободы). Идея ансамблевых методов состоит в том, чтобы уменьшить смещение и/или разброс таких слабых учеников путём объединения некоторых из них вместе, чтобы создать ***сильного ученика*** (или ***модель ансамбля***), который достигает лучших результатов [4].

Существуют несколько наиболее известных методов объединения базовых алгоритмов в композиции: голосование, взвешенное голосование, смесь экспертов (Mixture of Experts [32]). Данные методы часто применяются, когда базовые алгоритмы существенно отличаются друг от друга. В случаях, когда необходимо построить композицию используя один базовый алгоритм, широко применяется **бэггинг** (bagging, boostrap aggregation) [22]. Идея бэггинга состоит в том, что базовый алгоритм многократно обучается на случайных подвыборках с повторениями из обучающей выборки. Такой метод генерации подвыборок принято называть **бутсрап** (bootsrap). Похожим на бэггинг методом является **метод случайных подпространств** (random subspace method, RSM [27]). Его идея заключается в создании вариативности при обучении с помощью выбора случайных подмножеств признаков. Широко известным примером использования бэггинга и RSM является случайный лес [23]. 3 Другим известным способом ансамблирования базовых алгоритмов является **бустинг**. Идея бустинга состоит в жадном выборе очередного алгоритма для добавления в композицию так, чтобы он лучшим образом компенсировал имеющиеся на этом шаге ошибки. Широко известные примеры бустинга — AdaBoost [24] и градиентный бустинг (Gradient boosting [25]). **Стекинг** (Stacked generalization) был впервые предложен Д. Волпертом в 1992 году в работе [20] в достаточно общем виде. Основная идея стекинга заключается в использовании базовых классификаторов для получения предсказаний (метапризнаков) и дальнейшем использовании их как признаков для некоторого” обобщающего” алгоритма (иными словами, метаалгоритма). Иными словами, основной идеей стекинга является преобразование исходного пространства признаков задачи в новое пространство, точками которого являются предсказания базовых алгоритмов. Предлагается сначала выбрать набор пар произвольных подмножеств из обучающей выборки, затем для каждой пары обучить базовые алгоритмы на первом подмножестве и предсказать ими целевую переменную для второго подмножества. Предсказанные значения и становятся объектами нового пространства. В частности, автором рассматривается случай выбора всевозможнных пар подмножеств, в которых второе подмножество состоит из единственного объекта, а первое множество из всей обучающей выборки кроме этого объекта (leave-one-out). Очевидно, что такой способ позволяет перевести каждую точку исходного пространства признаков в точку нового пространства. Автор обобщает идею стекинга тем, что предлагает, обучая базовые классификаторы (первого уровня) над метапризнаками (первого уровня), получать метапризнаки второго уровня и так далее. Развитие идеи стекинга для задач многоклассовой классификации происходит в работах [30], [29]. В первой работе предлагается свести задачу мноклассовой классификации к набору задач одноклассовой классификации, которые решают базовые классификаторы, и обучения метаклассификатора над их предсказаниями. Во второй работе предлагается усложнить эту схему с помощью использования дополнительного уровня базовых классификаторов, построенных на предсказаниях базовых классификаторов первого уровня.

***Деревья решений*** являются популярными базовыми моделями для ансамблевых методов. Сильные ученики, состоящие из нескольких деревьев решений, можно назвать «лесами». Деревья, составляющие лес, могут быть выбраны либо неглубокими (глубиной в несколько узлов), либо глубокими (глубиной в множество узлов, если не в полную глубину со всеми листьями). Неглубокие деревья имеют меньший разброс, но более высокое смещение, и тогда для них лучшим выбором станут **последовательные методы**. Глубокие деревья, с другой стороны, имеют низкое смещение, но высокий разброс и, таким образом, являются подходящим выбором для бэггинга, который в основном направлен на уменьшение разброса [4].

**Случайный лес**представляет собой метод бэггинга, где глубокие деревья, обученные на бутстрап выборках, объединяются для получения результата с более низким разбросом. Тем не менее, случайные леса также используют другой прием, чтобы несколько обученных деревьев были менее коррелированными друг с другом: при построении каждого дерева вместо выбора всех признаков из датасета для генерации бутстрэпа выбирается и сохраняется только случайное их подмножество для построения дерева (обычно одинаковое для всех бутстрэп выборок) [4].

Выборка по признакам приводит к тому, что все деревья не смотрят на одну и ту же информацию для принятия своих решений и, таким образом, уменьшают корреляцию между различными возвращаемыми выходными данными. Другое преимущество выборки по признакам заключается в том, что **она делает процесс принятия решений более устойчивым к отсутствующим данным**: значения наблюдения (из обучающего датасета или нет) с отсутствующими данными можно восстанавливать с помощью регрессии или классификации на основе деревьев, которые учитывают только те признаки, где данные не отсутствуют. Таким образом, алгоритм случайного леса сочетает в себе концепции бэггинга и выбора подпространства случайных объектов для создания более устойчивых моделей [4].

## 1.3 Перестановочная важность

Перестановочная важность признаков — это метод проверки модели, который может быть использован для любой обученной модели, когда данные представимы в виде таблиц. Это особенно полезно для нелинейных или непрозрачных моделей. Перестановочная важность признаков определяется как уменьшение оценки точности модели машинного обучения при случайном перемешивании одного атрибута объекта. Такая процедура разрывает связь между признаком и целевым объектом, таким образом, снижение оценки модели указывает на то, насколько модель зависит от признака. Данный метод не зависит от модели и может быть вычислен много раз с различными перестановками признака [33].

Признаки, которые считаются малозначимыми для плохой модели (имеющие низкий процент перекрестной проверки), могут быть очень важны для хорошей модели. Поэтому всегда важно оценить прогностическую способность модели, используя проверенный набор (с перекрестной проверкой), прежде чем вычислять важность. Важность перестановки отражает не внутреннюю прогностическую ценность признака как такового, а то, насколько важен такой признак для конкретной модели.

Такая оценка может быть дана с помощью функции permutation\_importance библиотеки sklearn[33] - она вычисляет важность признаков для данного набора данных. Параметр n\_repeats задает количество случайного перемешивания признака.

Перестановочная важность признаков может быть вычислена либо на обучающем наборе, либо на наборе отложенного тестирования или проверки. Использование расширенного набора позволяет выделить, какие признаки вносят наибольший вклад в обобщающую способность проверяемой модели. Признаки, которые важны на тренировочном наборе, но не на расширенном наборе, могут привести к чрезмерной подгонке модели.

Важность функции перестановки оценивается как уменьшение оценки точности модели при случайном перемешивании одного признака. Функция оценки, которая будет использоваться для вычисления значимости, может быть задана с помощью аргумента scoring, который также принимает несколько оценщиков. Использование нескольких показателей более эффективно с точки зрения вычислений, чем последовательный вызов permutation\_importance несколько раз с другим показателем, поскольку он повторно использует прогнозы модели [33].

Ранжирование функций примерно одинаково для разных показателей, даже если шкалы значений важности сильно различаются. Однако, это не гарантируется, и разные метрики могут привести к значительному различию значимости признаков, в частности, для моделей, обученных для несбалансированных задач классификации, для которых выбор метрики классификации может иметь решающее значение.

Рассмотрим схему алгоритма вычисления важности перестановки:

* Входными данными являются: подобранная прогностическая модель , табличный набор данных (обучение или проверка) .
* Вычислить эталонную оценку модели по данным (например, точность для классификатора или для регрессора).
* Для каждого признака (столбца ):
  + Для каждого повторения в :
    - Произвольно перемешать столбец набора данных , чтобы сгенерировать измененную версию данных с именем .
    - Вычислить оценку модели по поврежденным данным .
  + Вычислить важность для признака , определенного по формуле 1.19:

Древовидные модели предоставляют альтернативную меру важности признаков, основанную на среднем уменьшении примесей (MDI). Примесь определяется количественно с помощью критерия разделения деревьев решений (критерий Джини, логарифмические потери [7] или среднеквадратичная ошибка). Однако данный метод может придавать большое значение признакам, которые могут не поддаваться прогнозированию на невидимых данных, когда модель переобучается. С другой стороны, важность признаков на основе перестановок позволяет избежать этой проблемы, поскольку она может быть вычислена на основе невидимых данных.

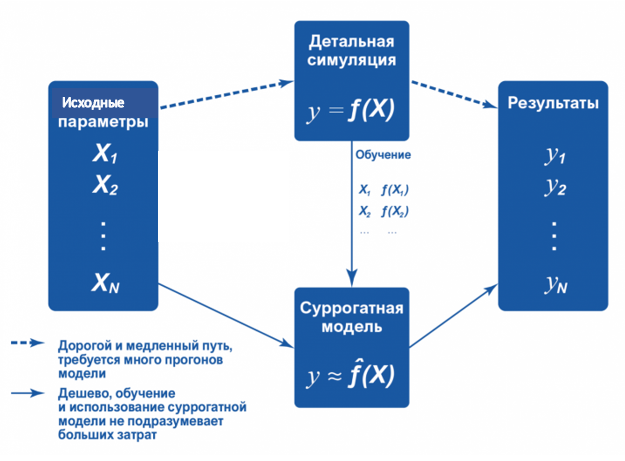
Кроме того, важность признаков на основе примесей для деревьев сильно смещена и отдает предпочтение признакам с высокой кардинальностью (обычно числовым объектам) по сравнению с объектами с низкой кардинальностью, такими как двоичные признаки или категориальные переменные с небольшим числом возможных категорий.

Значения функций, основанные на перестановках, не проявляют такого смещения. Кроме того, важность функции перестановки может быть вычисленной метрикой производительности для прогнозов модели и может использоваться для анализа любого класса модели (не только древовидных моделей).

Когда два признака коррелируются и один из признаков перемешивается, модель по-прежнему будет иметь доступ к признаку через его коррелированный объект. Это приведет к более низкому значению важности для обоих признаков, где они действительно могут быть важны. Один из способов справляться с таким поведением - группировать коррелированные признаки и сохранять только один признак из каждого кластера. [18]

## 1.4 Суррогатные модели

В задачах концептуального проектирования для экономии времени бывает необходимо заменить точную, но вычислительно сложную модель на быстро вычисляемую суррогатную модель. Суррогатная модель аппроксимирует зависимость, реализуемую рассматриваемой исходной моделью, с необходимой для последующего использования точностью, и строится по данным (конечному множеству пар “набор входных параметров” — “значения выходных характеристик модели для набора входных параметров”) [34].

Рисунок 1.2 Использование суррогатных моделей

Пусть задана некоторая обучающая выборка 𝑆𝑙𝑒𝑎𝑟𝑛. Кроме того, выберем некоторую технику построения базовых моделей 𝑔 из некоторого параметрического семейства 𝐺(под семейством могут пониматься модели типа Kriging, Artificial Neural Networks, Support Vector Regression, Multivariate Nonparametric Regression, Polynomial Regression и т.п.), которая позволяет по заданной выборке 𝑆𝑙𝑒𝑎𝑟𝑛 построить аппроксимацию 𝑦 = 𝑔(x, 𝜃), где 𝜃 — параметры модели. Под построением модели здесь понимается выбор для фиксированного семейства 𝐺 и выборки 𝑆𝑙𝑒𝑎𝑟𝑛-значений параметров 𝜃, которые задают некоторый элемент𝑔 = 𝑔(x, 𝜃) ∈ 𝐺, оптимальный в смысле некоторого критерия [34].

Предполагаемый процесс построения суррогатной модели состоит из следующих этапов.

1) Для того, чтобы избавиться от избыточности в признаках, проводится снижение размерности данных.

2) Декомпозиция пространства на области , которым соответствуют различные диапазоны значений выходов. Такая декомпозиция полезна, если требуется для разных диапазонов значений выходов обеспечить разную точность приближения.

3) Разбиение областей X𝑗 на связные подобласти X𝑗,𝑘 , соответствующие более регулярному поведению аппроксимируемой зависимости.

4) Построение по подвыборкам 𝑆𝑗,𝑘 = {(x, 𝑦) ∈ 𝑆𝑙𝑒𝑎𝑟𝑛 : x ∈ X𝑗,𝑘} соответствующих аппроксимирующих моделей.

5) Построение классификатора, который оценивает «близость» от данной точки x до подобластей X𝑗.

6) Построение итоговой суррогатной модели с помощью «сшивки» полученных аппроксимирующих моделей.

# ГЛАВА 2 МЕТОДЫ ОЦЕНКИ ЗНАЧИМОСТИ ПРИЗНАКОВ В ЗАДАЧЕ ОТБОРА ПРИЗНАКОВ

## 2.1 Задача отбора признаков

Целевые признаки (feature), которые используются для обучения модели, оказывают очень большое влияние на результаты. Неинформативные/слабо информативные признаки могут основательно понизить точность, производительность многих моделей, таких как линейная и логистическая регрессии. После устранения или преобразования неинформативных или слабо информативных признаков модель упрощается, и соответственно уменьшается размер набора данных в памяти и ускоряется работа алгоритмов машинного обучения на нем.

Можно выделить 3 задачи:

* **Выбор признаков (feature selection)** – подвыборка ненужных признаков (неинформативных, слабо информативных) и выбор признаков, имеющих наиболее большое значение взаимосвязи с выходной переменной.
* **Выделение и настройка признаков (feature extraction and feature engineering)** – превращение специфических данных предметной области в понятные для модели наборы чисел;
* **Трансформация признаков (feature transformation)** – преобразование данных для повышения точности/эффективности алгоритма (например, стандартизация данных).

Каждая из указанных задач направлена на обеспечение следующих преимуществ:

* **Уменьшение риска переобучения.** Чем меньше избыточных данных, тем меньше вероятность для модели принимать решения на основе «шума».
* **Повышение точности прогнозирования модели.** Чем меньше противоречивых данных, тем выше может быть точность.
* **Сокращение необходимого времени на обучение.** Чем меньше данных, тем быстрее может обучается модель.
* **Увеличение семантического понимания модели.**

 Выделяются методы ручного и автоматизированного отбора признаков.

Методы ручного отбора признаков основаны на анализе содержания каждого признака и/или их совокупности. Решение о включении либо исключении признака принимает исследователь.

В данном разделе будут рассмотрены некоторые методы автоматизированного отбора признаков (feature selection), которые применяются для подготовки данных. Такие методы уже могли быть реализованы с помощью Python и библиотеки scikit-learn. Подробное руководство по отбору признаков с помощью scikit-learn можно найти в документации к этой библиотеке в разделе Feature selection [35].

Методы отбора признаков делятся на три группы:

1) методы-фильтры (filter methods) оценивают признаки на основе информации из обучающей выборки. Такие методы применяются до запуска алгоритма обучения, на этапе предобработки;

2) методы-обертки (wrapper methods) - классификатор запускается на некоторым образом преобразованных множествах (в частности, подмножествах исходного множества) обучающей выборки, а затем некоторым образом выбирается множество наиболее полезных для обучения признаков (рисунок 2.1);



Рисунок 2.1 - Процесс работы оберточных методов

3) встроенные методы (embedded methods) не отделяют процессы отбора и обучения (рисунок 2.2). Методы данного типа также обладают рядом иных достоинств: хорошей приспособленностью к конкретной модели; отсутствием необходимости выделения специальных множеств для тестирования, как в предыдущих группах методов, и, как следствие из этого, меньший риск переобучения.



Рисунок 2.2 - Процесс работы встроенных методов

Рассмотрим подробнее каждую группу.

## 2.2 Методы – фильтры(filters)

Фильтрацией в широком смысле называется любое преобразование обрабатываемых переменных с целью изменения соотношения между их различными компонентами.

**Фильтры** (англ. *filter methods*) измеряют численное значение релевантности признаков на основе некоторой функции, и решают по определенному правилу, какие признаки стоит оставить в конечном множестве.

Фильтры классифицируются:

* Одномерные (англ. *univariate*) — функция определяет релевантность каждого признака по отношению к выходной переменной. В таком случае обычно измеряют некоторое условное "качество" каждого признака с помощью, например, статистических критериев (коэффициент ранговой корреляции Спирмена, критерий хи-квадрат и др.) и удаляют худшие. Например, библиотека scikit-learn содержит класс SelectKBest [16], реализующий одномерный отбор признаков (univariate feature selection) [35]. Указанный ранее класс можно применять совместно с различными статистическими критериями для отбора заданного количества признаков.
* Многомерные (англ. *multivariate*) — функция определяет релевантность некоторого подмножества исходного множества признаков относительно выходных меток.

*Критерий χ2 Пирсона* – это непараметрический метод, который позволяет оценить значимость различий между практическим (выявленным в результате исследования) количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое ожидается в изучаемых группах при справедливости гипотезы. Выражаясь проще, метод оценивает статистическую значимость различий нескольких относительных показателей (например, частот, долей) [36].

Критерий хи-квадрат был разработан и предложен в 1900 году английским математиком, статистиком, биологом и философом, основателем математической статистики и одним из основоположников биометрики Карлом Пирсоном(1857-1936).

Условия и ограничения применения критерия хи-квадрат Пирсона:

1. Сопоставляемые показатели должны быть измерены в номинальной шкале или в порядковой (например, оценки знаний учащихся, принимающая значения 2 и от 4 до 10).
2. Соответствующие группы должны быть независимыми, то есть критерий хи-квадрат не должен применяться при сравнении наблюдений "до-"после". В таких случаях проводится **тест** Мак-Немара(при сравнении двух связанных совокупностей) или рассчитывается Q-критерий Кохрена(в случае сравнения трех и более групп).

Рассмотрим алгоритм расчета критерия хи-квадрат Пирсона

1. Рассчитываем ожидаемое количество наблюденийдля каждой из ячеек таблицы сопряженности (при условии справедливости гипотезы об отсутствии взаимосвязи) путем перемножения сумм рядов и столбцов с последующим делением полученного произведения на общее число наблюдений.
2. Находим значение критерия χ2по формуле 2.1:

где i – номер строки (от 1 до r), j – номер столбца (от 1 до с), Oij – фактическое количество наблюдений в ячейке ij, Eij – ожидаемое число наблюдений в ячейке ij.

1. Определяем число степеней свободыпо формуле: f = (r – 1) × (c – 1). Соответственно, для четырехпольной таблицы, в которой 2 ряда (r = 2) и 2 столбца (c = 2), число степеней свободы составляет f2x2 = (2 - 1)\*(2 - 1) = 1.
2. Сравниваем значение критерия χ2 с критическим значением при числе степеней свободы f (по таблице).

В том случае, если полученное значение критерия χ2 больше критического, делаем вывод о наличии статистической взаимосвязи между изучаемым фактором риска и исходом при соответствующем уровне значимости [36].

*t-критерий Стьюдента*– общее название класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента. Наиболее частые случаи применения t-критерия связаны с проверкой равенства средних значений в двух выборках.

Данный критерий был разработан Уильямом Сили Госсетомдля оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны, статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент).

t-критерий Стьюдента используется для определения статистической значимости различий средних величин. Может применяться как в случаях сравнения независимых выборок (например, группы больных сахарным диабетом и группы здоровых), так и при сравнении связанных совокупностей (например, средняя частота пульса у одних и тех же пациентов до и после приема антиаритмического препарата). В последнем случае рассчитывается парный t-критерий Стьюдента [36].

Для применения t-критерия Стьюдента необходимо, чтобы исходные данные имели нормальное распределение. Также имеет значение равенство дисперсий (распределения) сравниваемых групп (гомоскедастичность). При неравных дисперсиях применяется t-критерий в модификации Уэлча (Welch's t) [36].

При отсутствии нормального распределения сравниваемых выборок вместо t-критерия Стьюдента используются аналогичные методы непараметрической статистики, среди которых наиболее известными является U-критерий Манна — Уитни [36].

Для сравнения средних величин t-критерий Стьюдента рассчитывается по формуле 2.2:

где **М1** - средняя арифметическая первой сравниваемой совокупности (группы), **М2** - средняя арифметическая второй сравниваемой совокупности (группы), **m1** - средняя ошибка первой средней арифметической, **m2** - средняя ошибка второй средней арифметической.

Полученное значение t-критерия Стьюдента необходимо правильно интерпретировать. Для этого нам необходимо знать количество исследуемых в каждой группе (n1 и n2). Находим число степеней свободы **f** по следующей формуле:

После этого определяем критическое значение t-критерия Стьюдента для требуемого уровня значимости (например, p=0,05) и при данном числе степеней свободы **f** по таблице [36].

При сравнении критического и рассчитанного значения критерия необходимо пользоваться следующими правилами:

* Если рассчитанное значение t-критерия Стьюдента равно или больше критического, найденного по таблице, делаем вывод о статистической значимости различий между сравниваемыми величинами.
* Если значение рассчитанного t-критерия Стьюдента меньше табличного, значит различия сравниваемых величин статистически не значимы.

*Парный t-критерий Стьюдента*– одна из модификаций метода Стьюдента, используемая для определения статистической значимости различий парных (повторных) измерений [36].

Парный t-критерий Стьюдента используется для сравнения двух зависимых (парных) выборок. Зависимыми являются измерения, выполненные у одних и тех же пациентов, но в разное время, например, артериальное давление у больных гипертонической болезнью до и после приема антигипертензивного препарата. Нулевая гипотеза гласит об отсутствии различий между сравниваемыми выборками, альтернативная - о наличии статистически значимых различий [36].

Основным условием является зависимость выборок, то есть сравниваемые значения должны быть получены при повторных измерениях одного параметра у одних и тех же пациентов.

Как и в случае сравнения независимых выборок, для применения парного t-критерия необходимо, чтобы исходные данные имели нормальное распределение. При несоблюдении этого условия для сравнения выборочных средних должны использоваться методы непараметрической статистики, такие как G-критерий знаков или Т-критерий Вилкоксона [36].

Парный t-критерий может использоваться только при сравнении двух выборок. Если необходимо сравнить три и более повторных измерений, следует использовать однофакторный дисперсионный анализ (ANOVA) для повторных измерений.

Парный t-критерий Стьюдента рассчитывается по формуле 2.4:

где Мd - средняя арифметическая разностей показателей, измеренных до и после, σd - среднее квадратическое отклонение разностей показателей, n - число исследуемых.

Интерпретация полученного значения парного t-критерия Стьюдента не отличается от оценки t-критерия для несвязанных совокупностей. Прежде всего, необходимо найти число степеней свободы f по формуле 2.5:

После этого определяем критическое значение t-критерия Стьюдента для требуемого уровня значимости (например, p<0,05) и при данном числе степеней свободы **f** по таблице [37].

*Критерий корреляции Пирсона* – это метод параметрической статистики, позволяющий определить наличие или отсутствие линейной связи между двумя количественными показателями, а также оценить ее тесноту и статистическую значимость. Другими словами, критерий корреляции Пирсона позволяет определить, изменяется ли (возрастает или уменьшается) один показатель в ответ на изменения другого? В статистических расчетах и выводах коэффициент корреляции обычно обозначается как rxy или Rxy [36].

Критерий корреляции Пирсона был разработан командой британских ученых во главе с Карлом Пирсоном (1857-1936) в 90-х годах 19-го века, для упрощения анализа ковариации двух случайных величин. Помимо Карла Пирсона над критерием корреляции Пирсона работали также Фрэнсис Эджуорт и Рафаэль Уэлдон.

Критерий корреляции Пирсона позволяет определить, какова теснота (или сила) корреляционной связи между двумя показателями, измеренными в количественной шкале. При помощи дополнительных расчетов можно также определить, насколько статистически значима выявленная связь [36].

Например, при помощи критерия корреляции Пирсона можно ответить на вопрос о наличии связи между температурой тела и содержанием лейкоцитов в крови при острых респираторных инфекциях, между ростом и весом пациента, между содержанием в питьевой воде фтора и заболеваемостью населения кариесом [36].

Условия и ограничения применения критерия хи-квадрат Пирсона**:**

1. Сопоставляемые показатели должны быть измерены в количественной шкале (например, частота сердечных сокращений, температура тела, содержание лейкоцитов в 1 мл крови, систолическое артериальное давление).
2. Посредством критерия корреляции Пирсона можно определить лишь наличие и силу линейной взаимосвязи между величинами. Прочие характеристики связи, в том числе направление (прямая или обратная), характер изменений (прямолинейный или криволинейный), а также наличие зависимости одной переменной от другой - определяются при помощи регрессионного анализа.
3. Количество сопоставляемых величин должно быть равно двум. В случае анализ взаимосвязи трех и более параметров следует воспользоваться методом факторного анализа.
4. Критерий корреляции Пирсона является параметрическим, в связи с чем условием его применения служит нормальное распределение каждой из сопоставляемых переменных. В случае необходимости корреляционного анализа показателей, распределение которых отличается от нормального, в том числе измеренных в порядковой шкале, следует использовать коэффициент ранговой корреляции Спирмена.
5. Следует четко различать понятия зависимости и корреляции. Зависимость величин обуславливает наличие корреляционной связи между ними, но не наоборот.

Расчет коэффициента корреляции Пирсона производится по следующей формуле:

Значения коэффициента корреляции Пирсона интерпретируются исходя из его абсолютных значений. Возможные значения коэффициента корреляции варьируют от 0 до ±1. Чем больше абсолютное значение rxy – тем выше теснота связи между двумя величинами. rxy = 0 говорит о полном отсутствии связи. rxy = 1 – свидетельствует о наличии абсолютной (функциональной) связи. Если значение критерия корреляции Пирсона оказалось больше 1 или меньше -1 – в расчетах допущена ошибка [36].

Для оценки тесноты, или силы, корреляционной связи обычно используют общепринятые критерии, согласно которым абсолютные значения rxy < 0.3 свидетельствуют о *слабой* связи, значения rxy от 0.3 до 0.7 - о связи *средней*тесноты, значения rxy > 0.7 - о *сильной*связи.

Более точную оценку силы корреляционной связи можно получить, если воспользоваться таблицей Чеддока(таблица 2.1):

Таблица 2.1 Оценка силы корреляционной связи

| Абсолютное значение rxy | Теснота (сила) корреляционной связи |
| --- | --- |
| менее 0.3 | слабая |
| от 0.3 до 0.5 | умеренная |
| от 0.5 до 0.7 | заметная |
| от 0.7 до 0.9 | высокая |
| более 0.9 | весьма высокая |

Оценка статистической значимости коэффициента корреляции rxy осуществляется при помощи t-критерия, рассчитываемого по формуле 2.7:

Полученное значение tr сравнивается с критическим значением при определенном уровне значимости и числе степеней свободы n-2. Если tr превышает tкрит, то делается вывод о статистической значимости выявленной корреляционной связи [36].

*Коэффициент ранговой корреляции Спирмена*– это непараметрический метод, который используется с целью статистического изучения связи между явлениями. В этом случае определяется фактическая степень параллелизма между двумя количественными рядами изучаемых признаков и дается оценка тесноты установленной связи с помощью количественно выраженного коэффициента [36].

Данный критерий был разработан и предложен для проведения корреляционного анализа в 1904 году Чарльзом Эдвардом Спирменом, английским психологом, профессором Лондонского и Честерфилдского университетов.

Коэффициент ранговой корреляции Спирмена используется для выявления и оценки тесноты связи между двумя рядами сопоставляемых количественных показателей. В том случае, если ранги показателей, упорядоченных по степени возрастания или убывания, в большинстве случаев совпадают (большему значению одного показателя соответствует большее значение другого показателя - например, при сопоставлении роста пациента и его массы тела), делается вывод о наличии прямой корреляционной связи. Если ранги показателей имеют противоположную направленность (большему значению одного показателя соответствует меньшее значение другого - например, при сопоставлении возраста и частоты сердечных сокращений), то говорят об **обратной** связи между показателями [37].

Коэффициент корреляции Спирмена обладает следующими свойствами:

1. Коэффициент корреляции может принимать значения от минус единицы до единицы, причем при rs=1 имеет место строго прямая связь, а при rs= -1 – строго обратная связь.
2. Если коэффициент корреляции отрицательный, то имеет место обратная связь, если положительный, то – прямая связь.
3. Если коэффициент корреляции равен нулю, то связь между величинами практически отсутствует.
4. Чем ближе модуль коэффициента корреляции к единице, тем более сильной является связь между измеряемыми величинами.

Условия использования коэффициента ранговой корреляции Спирмена:

В связи с тем, что коэффициент является методом непараметрического анализа, проверка на нормальность распределения не требуется.

Сопоставляемые показатели могут быть измерены как в непрерывной шкале, так и в порядковой (например, баллы экспертной оценки от 1 до 5).

Эффективность и качество оценки методом Спирмена снижается, если разница между различными значениями какой-либо из измеряемых величин достаточно велика. Не рекомендуется использовать коэффициент Спирмена, если имеет место неравномерное распределение значений измеряемой величины [36].

Расчет коэффициента ранговой корреляции Спирмена включает следующие этапы:

1. Сопоставить каждому из признаков их порядковый номер (ранг) по возрастанию или убыванию.
2. Определить разности рангов каждой пары сопоставляемых значений (*d*).
3. Возвести в квадрат каждую разность и суммировать полученные результаты.
4. Вычислить коэффициент корреляции рангов по формуле 2.8:
5. Определить статистическую значимость коэффициента при помощи t-критерия, рассчитанного по формуле 2.9

При использовании коэффициента ранговой корреляции условно оценивают тесноту связи между признаками, считая значения коэффициента меньше 0,3 - признаком слабой тесноты связи; значения более 0,3, но менее 0,7 - признаком умеренной тесноты связи, а значения 0,7 и более - признаком высокой тесноты связи.

Также для оценки тесноты связи может использоваться шкала Чеддока(таблица 2.1)

Статистическая значимость полученного коэффициента оценивается при помощи t-критерия Стьюдента. Если расчитанное значение t-критерия меньше табличного при заданном числе степеней свободы, статистическая значимость наблюдаемой взаимосвязи - отсутствует. Если больше, то корреляционная связь считается статистически значимой [36].

Преимуществом группы фильтров является низкая трудоёмкость вычисления релевантности признаков в наборе данных, но недостатком в таком подходе является игнорирование возможных зависимостей между признаками.

## 2.3 Методы-обертки (Wrapper methods)

Принцип методов обертки состоит в следующем [5]:

• выполняется поиск по пространству множеств исходного множества признаков;

• для каждого шага поиска используется информация о некотором качестве обучения на текущем множестве признаков (в качестве функции оценки качества обучения на текущей итерации используется точность классификатора).

Указанный процесс является циклическим и продолжается до тех пор, пока не будут достигнуты заданные условия останова. Оберточные методы учитывают зависимости между признаками, что является преимуществом по сравнению с фильтрами, к тому же показывают большую точность, но вычисления занимают длительное время, и повышается риск переобучения [5].

Существует несколько типов оберточных методов:

* детерминированные, которые изменяют множество признаков по определенному правилу
* рандомизированные, которые используют генетические алгоритмы для выбора искомого множества признаков.

Среди детерминированных алгоритмов самыми простыми являются алгоритмы последовательного поиска [5]:

* SFS (Sequential Forward Selection) — жадный алгоритм, который начинает с пустого множества признаков, на каждом шаге добавляя лучший из еще не выбранных признаков в результирующее множество. На первом этапе производительность классификатора оценивается по отношению к каждому признаку. Выбирается признак, который работает лучше всего. На втором этапе первый признак пробуется в сочетании со всеми другими функциями. Выбирается комбинация двух функций, обеспечивающих наилучшую производительность алгоритма. Процесс продолжается до тех пор, пока не будет выбрано указанное количество признаков.
* SBS (Sequential Backward Selection) — алгоритм обратный прямому, который начинает с изначального множества признаков, и удаляет по одному или несколько худших признаков на каждом шаге; следует отметить, что эмпирически обратный жадный алгоритм даёт обычно лучшие результаты по сравнению с прямым жадным алгоритмом. Это связано с тем, что обратный жадный алгоритм учитывает признаки, информативные в совокупности, но неинформативные, если рассматривать их по отдельности.
* SSS (Sequential Stepwise Selection) – двунаправленное исключение/отбор - комбинация вышеперечисленных алгоритмов: тестирование на каждом шаге после включения/исключения признаков.

На каждом шаге описанных алгоритмов для оценки качества обычно используются такие критерии, как F-тест (статистика Фишера), t-тест (статистика Стъюдента), скорректированный коэффициент детерминации R2 и прочие. Сам алгоритм при этом принимает форму последовательности F-тестов, t-тестов.

*U-критерий Манна-Уитни*– непараметрический статистический критерий, используемый для сравнения двух независимых выборок по уровню какого-либо признака, измеренного количественно. Метод основан на определении того, достаточно ли мала зона перекрещивающихся значений между двумя вариационными рядами (ранжированным рядом значений параметра в первой выборке и таким же во второй выборке). Чем меньше значение критерия, тем вероятнее, что различия между значениями параметра в выборках достоверны [36].

Данный метод выявления различий между выборками был предложен в 1945 году американским химиком и статистиком Фрэнком Уилкоксоном.  
В 1947 году он был существенно переработан и расширен математиками Х.Б. Манном(H.B. Mann) и Д.Р. Уитни(D.R. Whitney), по именам которых сегодня обычно и называется.

U-критерий Манна-Уитни используется для оценки различий между двумя независимыми выборками по уровню какого-либо количественного признака.U-критерий Манна-Уитни является непараметрическим критерием, поэтому, в отличие от t-критерия Стьюдента, не требует наличия нормального распределения сравниваемых совокупностей.U-критерий подходит для сравнения малых выборок: в каждой из выборок должно быть не менее 3 значений признака. Допускается, чтобы в одной выборке было 2 значения, но во второй тогда должно быть не менее пяти [36].

Условием для применения U-критерия Манна-Уитни является отсутствие в сравниваемых группах совпадающих значений признака (все числа – разные) или очень малое число таких совпадений.

Аналогом U-критерия Манна-Уитни для сравнения трех и более групп является Критерий Краскела-Уоллиса.

Рассмотрим схему расчета U-критерия Манна-Уитни [36]:

Сначала из обеих сравниваемых выборок составляется **единый ранжированный ряд**, путем расставления единиц наблюдения по степени возрастания признака и присвоения меньшему значению меньшего ранга. В случае равных значений признака у нескольких единиц каждой из них присваивается среднее арифметическое последовательных значений рангов.

В составленном едином ранжированном ряду общее количество рангов получится равным:

где n1 - количество элементов в первой выборке, а n2 - количество элементов во второй выборке.

Далее вновь разделяем единый ранжированный ряд на два, состоящие соответственно из единиц первой и второй выборок, запоминая при этом значения рангов для каждой единицы. Подсчитываем отдельно сумму рангов, пришедшихся на долю элементов первой выборки, и отдельно - на долю элементов второй выборки. Определяем большую из двух ранговых сумм (Tx) соответствующую выборке с nx элементами.

Наконец, находим значение U-критерия Манна-Уитни по формуле 2.10:

Полученное значение U-критерия сравниваем по таблице для избранного уровня статистической значимости (p=0.05 или p=0.01) с критическим значением U при заданной численности сопоставляемых выборок:

* Если полученное значение U меньшетабличного или равноему, то признается статистическая значимость различий между уровнями признака в рассматриваемых выборках (принимается альтернативная гипотеза). Достоверность различий тем выше, чем меньше значение U.
* Если же полученное значение U большетабличного, принимается нулевая гипотеза.

*Критерий Уилкоксона для связанных выборок*(также используются названия Т-критерий Уилкоксона, критерий Вилкоксона, критерий знаковых рангов Уилкоксона, критерий суммы рангов Уилкоксона) – непараметрический статистический критерий, используемый для сравнения двух связанных (парных) выборок по уровню какого-либо количественного признака, измеренного в непрерывной или в порядковой шкале [36].

Суть метода состоит в том, что сопоставляются абсолютные величины выраженности сдвигов в том или ином направлении. Для этого сначала все абсолютные величины сдвигов ранжируются, а потом суммируются ранги. Если сдвиги в ту или иную сторону происходят случайно, то и суммы их рангов окажутся примерно равны. Если же интенсивность сдвигов в одну сторону больше, то сумма рангов абсолютных значений сдвигов в противоположную сторону будет значительно ниже, чем это могло бы быть при случайных изменениях [36].

Тест был впервые предложен в 1945 году американским статистиком и химиком Фрэнком Уилкоксоном (1892-1965). В той же научной работе автором был описан еще один критерий, применяемый в случае сравнения независимых выборок.

Т-критерий Уилкоксона используется для оценки различий между двумя рядами измерений, выполненных для одной и той же совокупности исследуемых, но в разных условиях или в разное время. Данный тест способен выявить направленность и выраженность изменений - то есть, являются ли показатели больше сдвинутыми в одном направлении, чем в другом [36].

Классическим примером ситуации, в которой может применяться Т-критерий Уилкоксона для связанных совокупностей, является исследование "до-после", когда сравниваются показатели до и после лечения. Например, при изучении эффективности антигипертензивного средства сравнивается артериальное давление до приема препарата и после приема.

Условия и ограничения применения Т-критерия Уилкоксона:

1. Критерий Уилкоксона является непараметрическим критерием, поэтому, в отличие от парного t-критерия Стьюдента, не требует наличия нормального распределения сравниваемых совокупностей.
2. Число исследуемых при использовании T-критерия Уилкоксона должно быть не менее 5.
3. Изучаемый признак может быть измерен как в количественной непрерывной (артериальное давление, ЧСС, содержание лейкоцитов в 1 мл крови), так и в порядковой шкале (число баллов, степень тяжести заболевания, степень обсемененности микроорганизмами).
4. Данный критерий используется только в случае сравнения двух рядов измерений. Аналогом Т-критерия Уилкоксона для сравнения трех и более связанных совокупностей является Критерий Фридмана.

Рассмотрим схему рассчета Т-критерия Уилкоксона для связанных выборок.

1. Вычислить разность между значениями парных измерений для каждого исследуемого. Нулевые сдвиги далее не учитываются.
2. Определить, какие из разностей являются типичными, то есть соответствуют преобладающему по частоте направлению изменения показателя.
3. Проранжировать разности пар по их абсолютным значениям (то есть, без учета знака), в порядке возрастания. Меньшему абсолютному значению разности приписывается меньший ранг.
4. Рассчитать сумму рангов, соответствующих нетипичным сдвигам.

Таким образом, Т-критерий Уилкоксона для связанных выборок рассчитывается по формуле 2.11:

где ΣRr - сумма рангов, соответствующих нетипичным изменениям показателя.

Полученное значение T-критерия Уилкоксона сравниваем с критическим по таблице для избранного уровня статистической значимости (p=0.05или p=0.01) при заданной численности сопоставляемых выборок n [37]:

* Если расчетное (эмпирическое) значение Тэмп. меньше табличного Ткр. или равно ему, то признается статистическая значимость изменений показателя в типичную сторону (принимается альтернативная гипотеза). Достоверность различий тем выше, чем меньше значение Т.
* Если Тэмп. больше Ткр., принимается нулевая гипотеза об отсутствии статистической значимости изменений показателя.

Популярным оберточным методом также является SVM-RFE (SVM-based Recursive Feature Elimination, рекурсивное исключение признаков), который иногда также относят к встроенным методам отбора. Данный метод использует как классификатор SVM и работает итеративно: начиная с полного множества признаков обучает классификатор, ранжирует признаки по весам, которые им присвоил классификатор, убирает какое-то число признаков и повторяет процесс с оставшегося подмножества признаков, если не было достигнуто их требуемое количество. Таким образом, указанный метод очень похож на встроенный, потому что непосредственно использует знание того, как устроен классификатор [5].

## 2.4 Встроенные методы (embedded)

Встроенные алгоритмы требуют меньше вычислений, чем обёрточные методы (хотя и больше, чем методы фильтрации).

Они также похожи на оберточные методы, однако для выбора признаков используется структура некоторой модели машинного обучения. В оберточных методах классификатор служит только для оценки работы на данном множестве признаков, тогда как встроенные методы используют некоторую информацию о признаках, которую классификаторы присваивают во время обучения.

Основным методом из этой категории является **регуляризация** [3].

Регуляризация — это штраф за сложность модели, который защищает от переобучения в случае наличия слабо информативных признаков. Ошибочным будет и мнение, что регуляризация бывает только в линейных моделях - и для бустинга, и для нейросетей существуют свои методы регуляризации. Пример регуляризации для бинарных деревьев описан в пункте 1.1.

Существуют различные разновидности регуляризации, но основной принцип общий. Если рассмотреть работу классификатора без регуляризации, то она состоит в построении такой модели, которая наилучшим образом настроилась бы на предсказание всех точек тренировочного сета.  
Например, если алгоритм классификации линейная регрессия, то подбираются коэффициенты полинома, который аппроксимирует зависимость между признаками и целевой переменной. В качестве оценки качества подобранных коэффициентов выступает среднеквадратичная ошибка (RMSE) [12]. Т.е. параметры подбираются так, чтобы суммарное отклонение (точнее суммарный квадрат отклонений) у точек, предсказанных классификатором от реальных точек, было минимальным. Идея регуляризации в том, чтобы построить алгоритм, минимизирующий не только ошибку, но и количество используемых переменных.

Данная группа методов часто применяются для борьбы с переизбыточностью данных, когда независимые переменные коррелируют друг с другом (т.е. имеет место мультиколлинеарность). Следствием этого является плохая обусловленность матрицы  и неустойчивость оценок коэффициентов регрессии. Оценки, например, могут иметь неправильный знак или значения, которые намного превосходят те, которые приемлемы из физических или практических соображений.

Обе техники позволяют уменьшить размерность данных и устранить/смягчить проблему переобучения. Для этого применяются два способа:

* L1-регуляризация — добавляет штраф к сумме абсолютных значений коэффициентов (формула 2.12). Данный метод используется в **Лассо-регрессии**. В процессе работы алгоритма величина приписанных алгоритмом коэффициентов будет пропорциональна важности соответствующих переменных для классификации, а для переменных, которые дают наименьший вклад в устранение ошибки, коэффициенты станут нулевыми. Таким образом, более значимые признаки сохранят свои коэффициенты ненулевыми, а менее значимые – обнулятся. Стоит также отметить, что большие по модулю отрицательные значения коэффициентов тоже говорят о сильном влиянии.
* L2-регуляризация — добавляет штраф к сумме квадратов коэффициентов (формула 2.13). Указанный выше метод используется в **ридж**-регрессии.

Уделим больше внимания параметру альфа. Он позволяет настраивать вклад регуляризирующего оператора в общую сумму. С его помощью указывается приоритет — точность модели или минимальное количество используемых переменных.  
Несмотря на незначительность различия, свойства отличаются. Если в ridge по мере роста альфа все коэффициенты получают значения все ближе к нулевым, но обычно при этом все-таки не зануляются, то в LASSO с ростом альфа все больше коэффициентов становятся нулевыми и совсем перестают вносить вклад в модель. Более значимые признаки сохранят свои коэффициенты ненулевыми, менее значимые — обнулятся [2,3,12].

**Методы преобразования признакового пространства** некоторым образом составляют из исходных признаков новые, все также полностью описывающие пространство набора данных, но уменьшая его размерность и теряя в репрезентативности данных, т.к. становится непонятно, за что отвечают новые признаки. Все методы feature extraction можно разделить на **линейные** и **нелинейные**.

Одним из самых известных методов **линейного** выделения признаков является PCA (Principal Component Analysis, рус. *метод главных компонент*). Основной идеей этого метода является поиск такой гиперплоскости, на которую при ортогональной проекции всех признаков максимизируется дисперсия. Данное преобразование может быть произведено с помощью сингулярного разложения матриц и создает проекцию только на линейные многомерные плоскости, поэтому и метод находится в категории линейных [5].

# ГЛАВА 3 ПРИМЕНЕНИЕ МЕТОДОВ ОЦЕНКИ ВАЖНОСТИ ПРИЗНАКОВ В РЕАЛЬНОЙ ЗАДАЧЕ

## 3.1 Описание задачи

Рассмотрим применение методов-фильтров и встроенных методов на датасете для задачи [8] – предсказать, зарабатывает ли человек больше $50 тыс. Загрузим библиотеки и данные(число объектов в датасете – 32562, вместо текстовых значений целевой переменной используем метки 0 и 1: ‘ <=50K’ = 0, ’>50K’ = 1), для удобства оставив только численные признаки, такие как:

* age – возраст
* fnlwgt (final weight) – примерная оценка количества людей, которое представляет каждая строка данных
* educational-num – длительность обучения
* capital-gain – прирост капитала
* capital-loss – потеря капитала
* hours-per-week – количество рабочих часов в неделю

## 3.2 Обучение моделей

Все кросс-валидации – 5-кратные перекрестные с сохранением приблизительной частоты появления классов в выборках и перемешиванием с жестко заготовленным случайным распределением для повторяемости [13]. Используемые далее обозначения: train\_scores – оценки точности классификатора на каждой из тренировочных выборок, test\_scores – на каждой из тестовых выборок, mean\_score – математическое ожидание предыдущей величины. Для стабилизации дисперсии будем использовать класс-препроцессор данных PowerTransformer [14] по методу Йео-Джонсона (Yeo-Johnson) [15]. Коэффициенты линейных моделей нормируются на их сумму.

Оценка точности классификатора на кросс-валидации для случайного леса:

train scores = [0.99998025 0.99997231 0.99997693 0.9999678 0.99997414]

mean score = 0.99997 +/- 0.00000

test score = [0.82427915 0.82290796 0.83106668 0.8192637 0.83155106]

mean score = 0.82581 +/- 0.00478

Важность признаков для случайного леса (рисунок 3.1):

Рисунок 3.1. – Важность признаков случайного леса

Самым важным признаком для случайного леса [10] является fnlwgt. Это можно интерпретировать как то, что главным фактором того, что человек зарабатывает больше $50 тыс. является количество людей с такими же характеристиками. Происходит это потому, что деревья решений и случайные леса могут выдавать сильно смещённую оценку признаков [6]. Притом, чем выше риск переобучения модели, тем сильнее может быть смещение, поэтому доверять оценкам таких моделей надо с осторожностью. В данном случае видно, что модель переобучена (математическое ожидание точности классификатора случайного леса на тестовой выборке составило 99%, а на тестовой – всего 82%) и выдает смещенные оценки.

Рассмотрим оценку точности с помощью суррогатных моделей:

Оценка точности классификатора на кросс-валидации для суррогата случайного леса:

train scores = [0.99997586 0.99998316 0.9999789 0.99997561 0.99997187]

mean score = 0.99998 +/- 0.00000

test score = [0.86122731 0.86878913 0.87685501 0.86725615 0.86313271]

mean score = 0.86745 +/- 0.00543

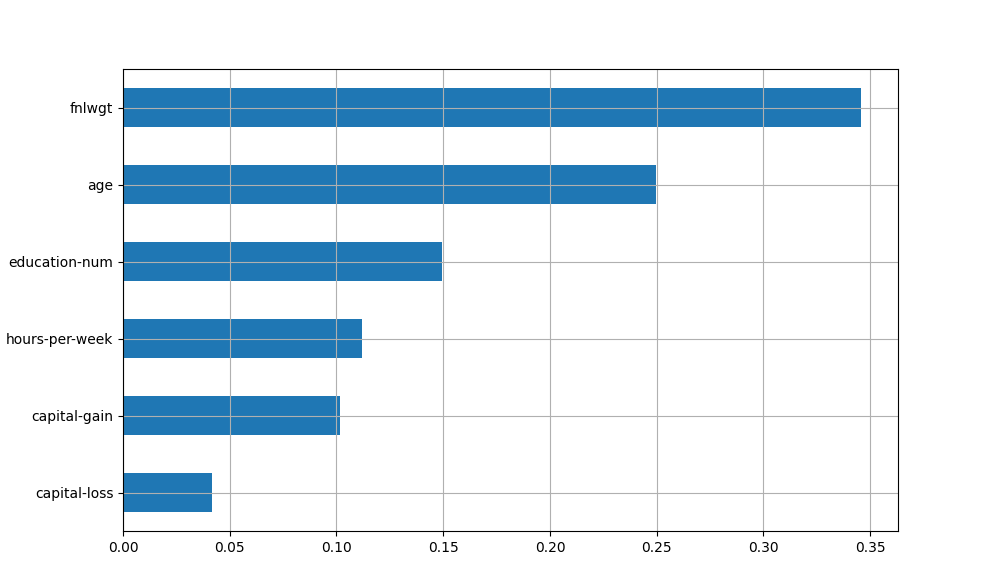
Важность признаков для суррогата случайного леса (рисунок 3.2):

Рисунок 3.2 Оценка важности признаков суррогатной моделью случайного леса.

Заметим, что оценка параметра fnlwgt упала, а практически всех остальных

– выросла, хотя их порядок и был сохранен. Это показывает хорошую согласованность оценок суррогатных моделей с оценками исходных моделей.

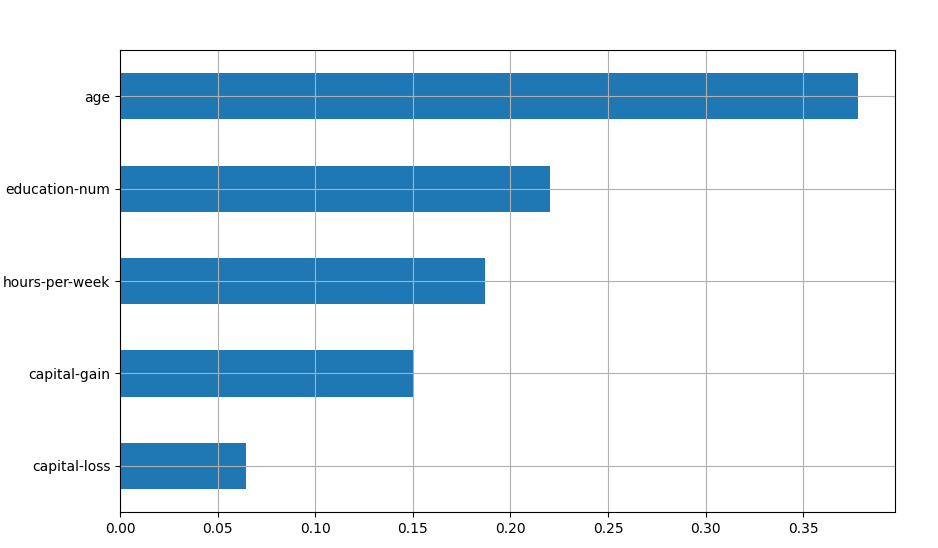
Методом прямого последовательного отбора признаков отберем наилучшее подмножество из 5 признаков (рисунок 3.3):

Рисунок 3.3. Важность признаков для метода прямого последовательного отбора признаков для случайного леса

Заметим, что наименее важным признаков оказался fnlwgt, а остальные признаки сохранили отношение порядка.

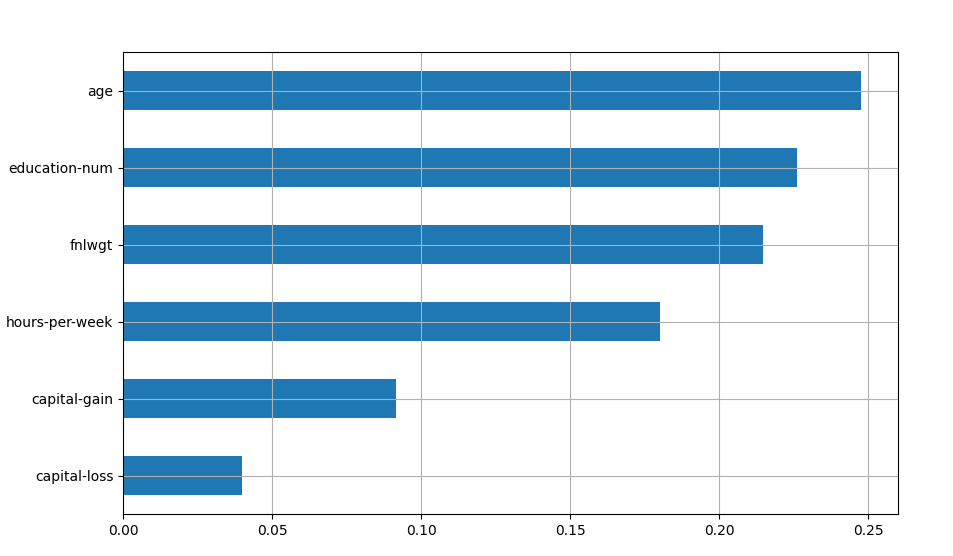
Оценим важность признаков методом перестановочной важности (рисунок 3.4):

Рисунок 3.4 Оценка важности признаков методом перестановочной важности

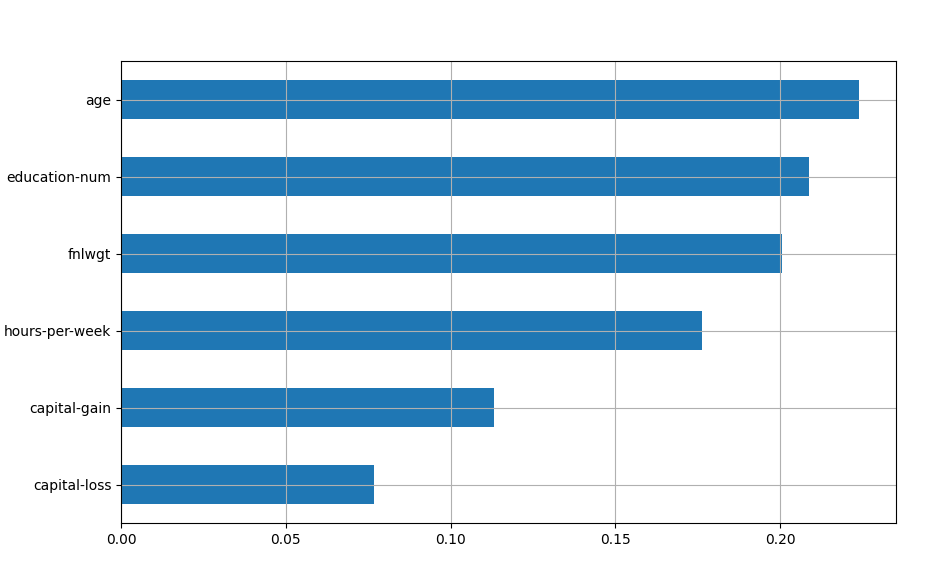
Отметим, что параметр fnlwgt получил значительно меньшую оценку, чем при оценке случайным лесом, хоть и остался в тройке лидирующих по важности. Оставшиеся признаки также показывают сохранение отношения порядка, полученных иными методами.

Далее опишем предлагаемую модификацию метода определения важности признаков, основанную на определении важности с помощью перестановочной важности, рассмотренной ранее. В частности, вот так была изменена формула подсчета важности каждого отдельного признака:

( -> )

Рассмотрим применение данного метода на практике (рисунок 3.5).

В отличие от обычной перестановочной важности признаков, её модифицированная версия склонна к увеличению важности параметров, которые были оценены низко другими метриками. Так же было сохранено отношение порядка, полученного оригинальным методом, что показывает согласованность данной разработки с показанными ранее методами.

Рисунок 3.5 Важность признаков, оценённых модифицированным методом

Точность на кросс-валидации и коэффициенты регрессии для логистических регрессий:

Повторим процедуру для линейной модели (с L1-регуляризацией).

Точность на кросс-валидации для линейной модели с L1-регуляризацией:

train scores = [0.82988457 0.82727422 0.82610092 0.83029481 0.82600082]

mean score = 0.82791 +/- 0.00184

test score = [0.82034993 0.83000963 0.8348707 0.81787667 0.83548066]

mean score = 0.82772 +/- 0.00732

Для линейной модели с L2-регуляризацией:

Точность на кросс-валидации для линейной модели с L2-регуляризацией:

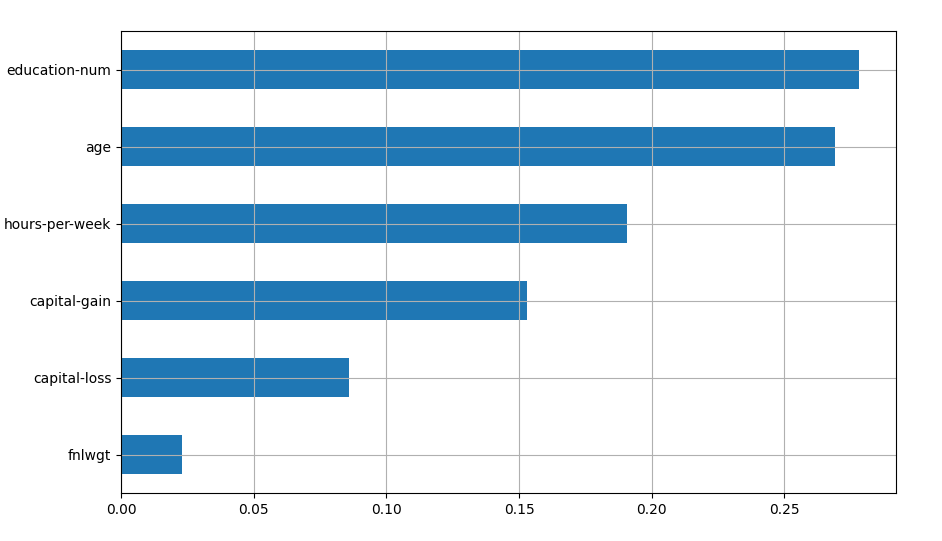
train scores = [0.82988494 0.82727349 0.82610074 0.83029575 0.82600163]

mean score = 0.82791 +/- 0.00184

test score = [0.82034877 0.83001208 0.83487096 0.81787925 0.83548092]

mean score = 0.82772 +/- 0.00732

Коэффициенты регрессии для линейной модели с L1-регуляризацией (рисунок 3.6):

Рисунок 3.6. – Коэффициенты регрессии логистической регрессии с L1-регуляризацией

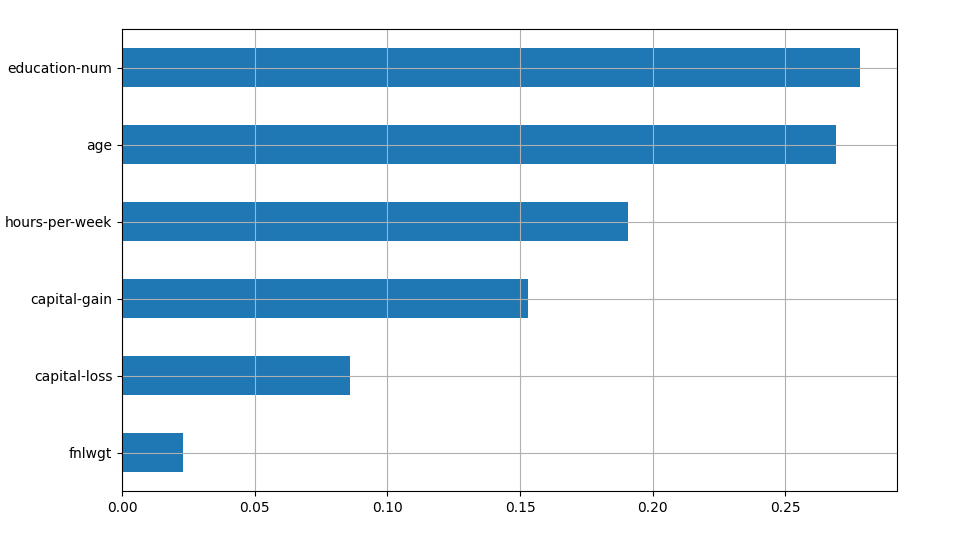
Коэффициенты регрессии для линейной модели с L2-регуляризацией (рисунок 3.7):

Рисунок 3.7. – Коэффициенты регрессии логистической регрессии с L2-регуляризацией

В отличие от случайного леса, регуляризация позволила сразу определить низкую важность параметра fnlwgt, что хорошо согласуется с выбором метода прямого последовательного отбора признаков. Отметим также, что модели оказались не переобучены (точности классификатора как на тестовой, так и на тренировочной выборках оказались равными 82,8%), а параметр education-num получил оценку важности, сходную с оценкой параметра age.

## 3.3 Внесение дополнительных шумовых признаков

Создадим 12 шумовых признаков, элементами которых будут некоррелируемые случайные числа из выборок с нормальным, равномерным и Лапласовым распределениями. Параметры каждого распределения подбираются случайным образом независимо друг от друга (рисунок 3.8).



Рисунок 3.8. - Графики распределения плотности вероятностей шумовых признаков

Покажем, что шумовые признаки слабо коррелируют с целевой переменной с помощью функции взаимной информации (mutual information) (таблица 3.1):

Таблица 3.1. Взаимная информация признаков по сравнению с целевой переменной

|  |  |
| --- | --- |
| feature\_name | Mutual information |
| capital-gain | 0.080221 |
| age | 0.065703 |
| education-num | 0.064743 |
| hours-per-week | 0.043655 |
| capital-loss | 0.033617 |
| fnlwgt | 0.033390 |
| norm\_3 | 0.003217 |
| unif\_3 | 0.002696 |
| norm\_0 | 0.002506 |
| norm\_2 | 0.002052 |
| lapl\_3 | 0.001201 |
| unif\_1 | 0.001144 |

Заметим, что основные признаки имеют на порядок большую оценку количества взаимной информации по сравнению с шумовыми.

Точность на кросс-валидации для случайного леса с внесенными шумовыми признаками:

train scores = [1. 1. 1. 1. 1.]

mean score = 1.00000 +/- 0.00000

test score = [0.8522425 0.85382173 0.86249657 0.84897581 0.85443027]

mean score = 0.85439 +/- 0.00447

Важность признаков для случайного леса (рисунок 3.8):

Рисунок 3.8. – Важность признаков случайного леса с шумовыми признаками

Несмотря на большое количество добавленных шумовых признаков, точность модели на кросс-валидации возросла как на каждой проверке, так и в среднем! Кроме этого, все шумовые признаки имеют высокую важность, сравнимую с двумя оригинальными. Очевидно, что наша модель переобучена (оценка математического ожидания точности классификатора не только достигла единицы на тренировочной выборке, но и на тестовой осталась на уровне 85%), однако в реальных задачах такие ситуации бывает очень сложно распознать, особенно когда при удалении некоторых признаков (про которые неизвестно – шумовые они, или нет) падает валидационная точность. Кроме того, часто бывает сложно подобрать пороговое значение важности признаков для исключения их из модели.

Посмотрим, как изменились оценки важности признаков для суррогата случайного леса.

Точность классификатора для суррогата случайного леса на зашумленных данных:

train scores = [1. 1. 1. 1. 1.]

mean score = 1.00000 +/- 0.00000

test score = [0.86590031 0.8665263 0.86342334 0.87133953 0.8588441]

mean score = 0.86521 +/- 0.00409

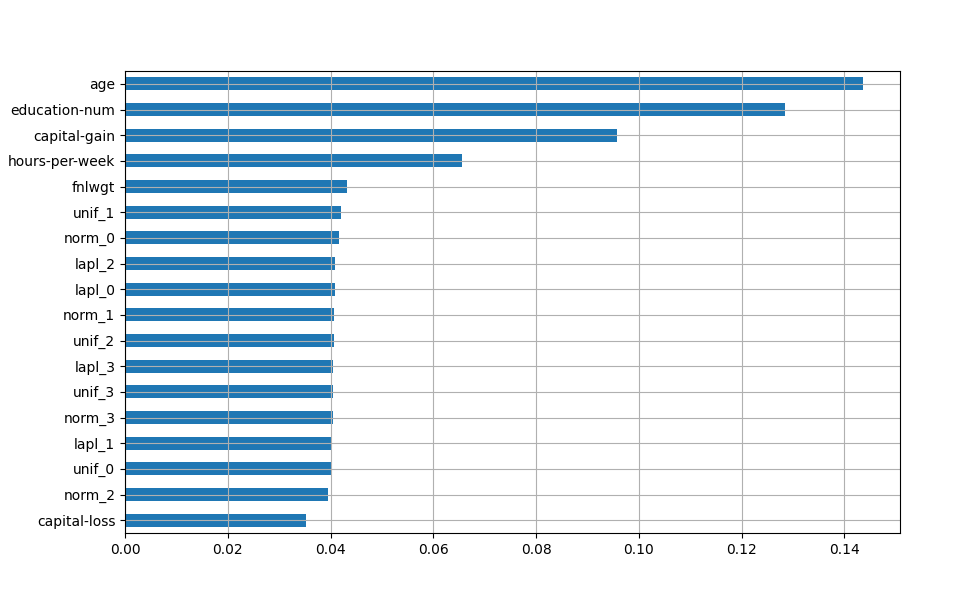
Оценка важности признаков для суррогата случайного леса (рисунок 3.9):

Рисунок 3.9 Важность признаков суррогата случайного леса на зашумленных данных

Характеристика education-num значительно выросла вместе с capital-gain, а вот оценка параметра age упала по сравнению с оригинальной. Отношение порядка сохранено.

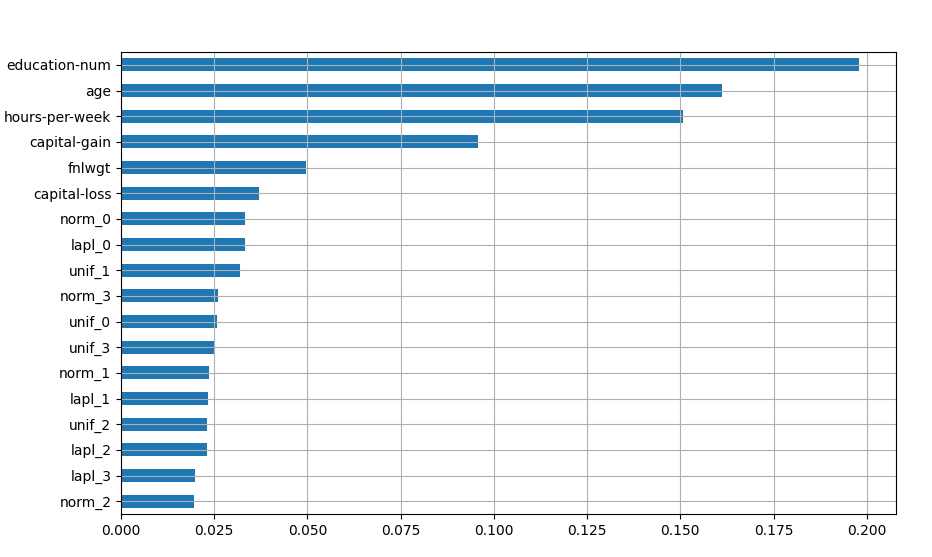
Оценка важности признаков методом перестановки (рисунок 3.10):

Рисунок 3.10 Оценки важности признаков методом перестановок

Шумовые признаки оценены как менее важные по сравнению с оригинальными.

Рассмотрим влияние разработанной модификации на оценки важности.

Оценка важности признаков модифицированным методом перестановки (рисунок 3.11)

Оценки шумовых признаков достигли значений, соответствующих значениям в случайном лесу, а признаки сохранили порядок, оценённый оригинальным методом.

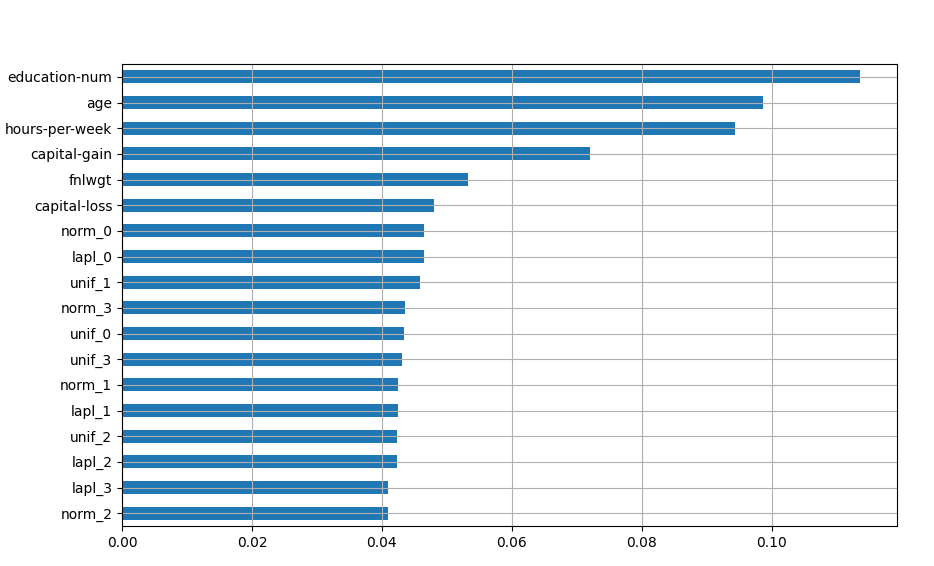
Точность на кросс-валидации при зашумлённых данных для линейной модели с L1-регуляризацией:

train scores = [0.83019323 0.82751386 0.82642378 0.83058806 0.82619654]

mean score = 0.82818 +/- 0.00186

test score = [0.81993058 0.83005516 0.83446553 0.81763029 0.83543145]

mean score = 0.82750 +/- 0.00738

Рисунок 3.11 Оценки важности признаков модифицированным перестановочным методом.

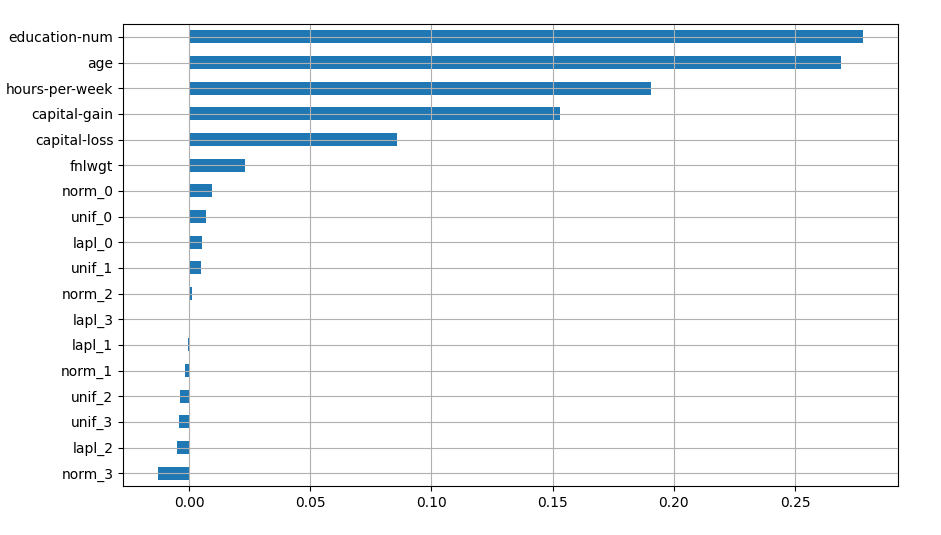
Коэффициенты регрессии линейной модели с L1-регуляризацией (рисунок 3.12):

Рисунок 3.12. – Коэффициенты регрессии с шумовыми признаками при L1-регуляризации

Точность на кросс-валидации при зашумлённых данных линейной модели с L2-регуляризацией:

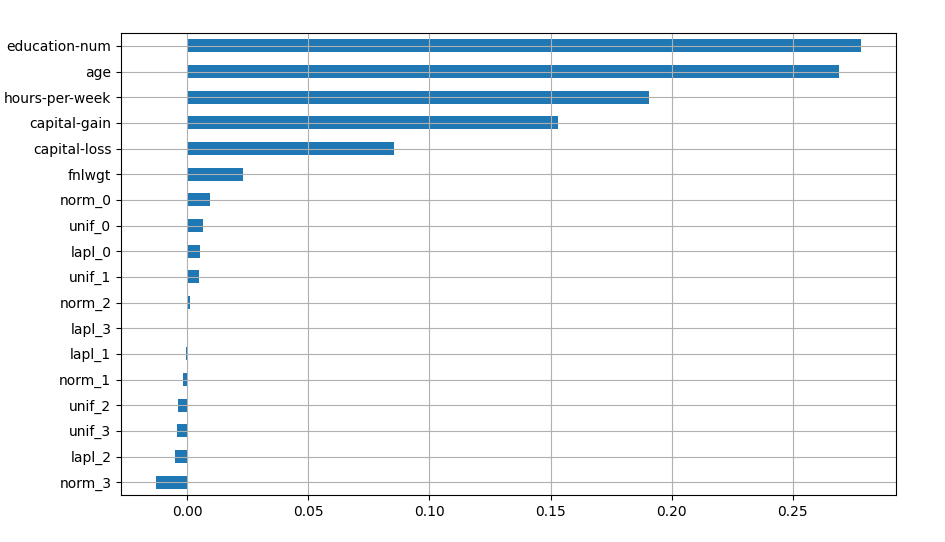
train scores = [0.83019479 0.827513 0.82642391 0.83058941 0.8261957]

mean score = 0.82818 +/- 0.00186

test score = [0.8199213 0.83005284 0.83446514 0.817625 0.8354299]

mean score = 0.82750 +/- 0.00739

После добавления шумовых признаков модель не переобучилась, к тому же указанные выше признаки имеют значительно меньшие коэффициенты, чем оригинальные. Отметим, что распределение коэффициентов в линейных моделях часто зависит от способа нормализации или масштабирования признаков (рисунок 3.13).

Рисунок 3.13. – Коэффициенты регрессии с шумовыми признаками при L2-регуляризации

## 3.4 Подбор гиперпараметров

Проведём отбор признаков статистическими методами, для чего будем использовать обобщённый вариант SelectKBest и SelectPercentile, который называется GenericUnivariateSelect [16]. Он принимает на вход 3 параметра – функцию оценки, режим отбора и его характеристики. В качестве функции оценки будем использовать взаимную информацию.

Сгенерированные нами признаки имеют низкое значение оценочной функции (scores\_), поэтому в дальнейшем селектор не будет их использовать (get\_support() =False)[16].

В реальной задаче (когда количество шумовых признаков неизвестно) параметры GenericUnivariateSelect можно находить на кросс-валидации вместе с другими гиперпараметрами [17] модели. Посмотрим, как изменится точность классификаторов после подбора их гиперпараметров, а также количества признаков селектора:

Точность на кросс-валидации для случайного леса с наилучшими гиперпараметрами модели (best\_params):

train scores = [0.91956644 0.91766083 0.91975952 0.9197495 0.91516779]

mean score = 0.91838 +/- 0.00179

test score = [0.84414441 0.84622556 0.8541299 0.84239658 0.84981241]

mean score = 0.84734 +/- 0.00420

best params = {'rf\_max\_depth': 12, 'rf\_\_max\_features': 0.1, 'selector\_param': 5}

Точность на кросс-валидации значительно выросла, а лучший результат получился всего для 5 признаков (рисунок 3.14):



Рисунок 3.14. – Важность признаков случайного леса после фильтрации

Такой результат был получен после удаления шумовых признаков и признака fnlwgt, который при первоначальной оценке был самым значимым для модели. Однако из всех оригинальных признаков он имел наименьшее значение оценочной функции как в GenericUnivariateSelect, так и в методе прямого последовательного отбора признаков, так и для линейных моделей с регуляризацией. Результаты оценки важности признаков после их отбора и настройки модели имеют более логичную интерпретацию – на заработок человека влияют именно характеристики человека, а не параметры самой выборки. Таким образом, статистический отбор признаков бывает полезен для увеличения точности некоторых типов моделей и получения менее смещённой оценки при интерпретации их результатов.

Рассмотрим суррогатную модель случайного леса после подбора наилучших значений гиперпараметров модели.

Точности классификатора на кросс-валидации для суррогата случайного леса после подбора гиперпараметров (best\_params):

best params = {'rf\_\_max\_depth': 12, 'rf\_\_max\_features': 0.1, 'selector\_\_param': 5}

train scores = [0.91185327 0.91131239 0.91662501 0.91089456 0.91029048]

mean score = 0.91220 +/- 0.00227

test score = [0.89389047 0.89511106 0.86838463 0.89805705 0.89928409]

mean score = 0.89095 +/- 0.01145

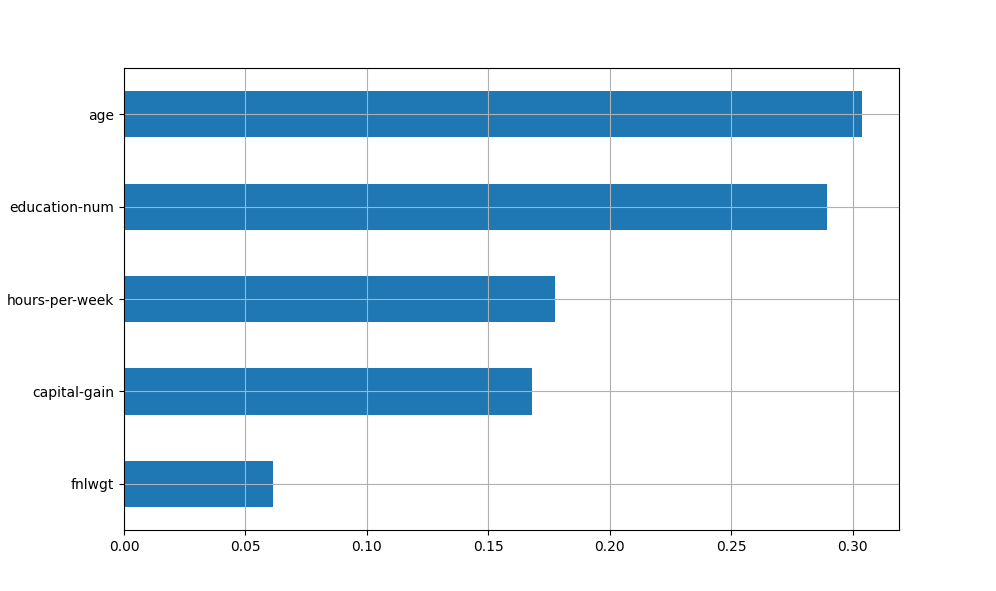
Оценки важности признаков суррогата случайного леса после подбора гиперпараметров модели (рисунок 3.15

Рисунок 3.15 Оценки важности признаков суррогатной моделью случайного леса после подбора гиперпараметров модели.

Как можно заметить из таблицы 1, параметры fnlwgt и capital-loss имеют близкие показатели оценки взаимной информации, что показывает высокую вероятность подмены одного из них другим при оценках важности. То же самое и с оценками параметров age и education-num.

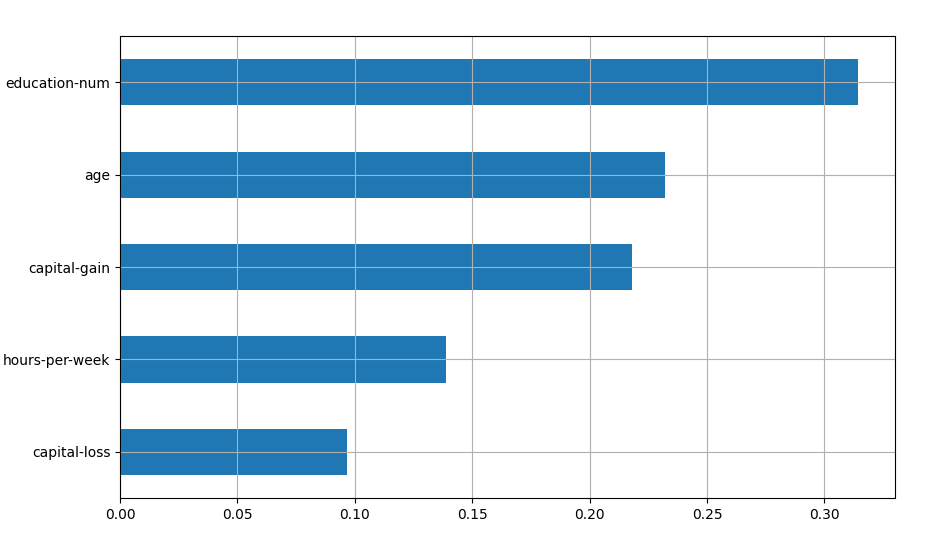
Оценка важности признаков методом перестановки после подбора гиперпараметров модели (рисунок 3.16):

Рисунок 3.16 Оценки важности признаков методом перестановки после подбора гиперпараметров модели.

Оценка важности признаков методом перестановки после подбора гиперпараметров модели (рисунок 3.17):

Заметно, что признаки находятся в том же отношении порядка, что и при применении не модифицированного метода. Хоть оригинальный метод и расходится с оценкой важности случайного леса, такое же поведение проявляет и модифицированный метод.

Посмотрим, как изменится точность классификатора после подбора коэффициента регуляризации у логистической регрессии L1-регуляризации.

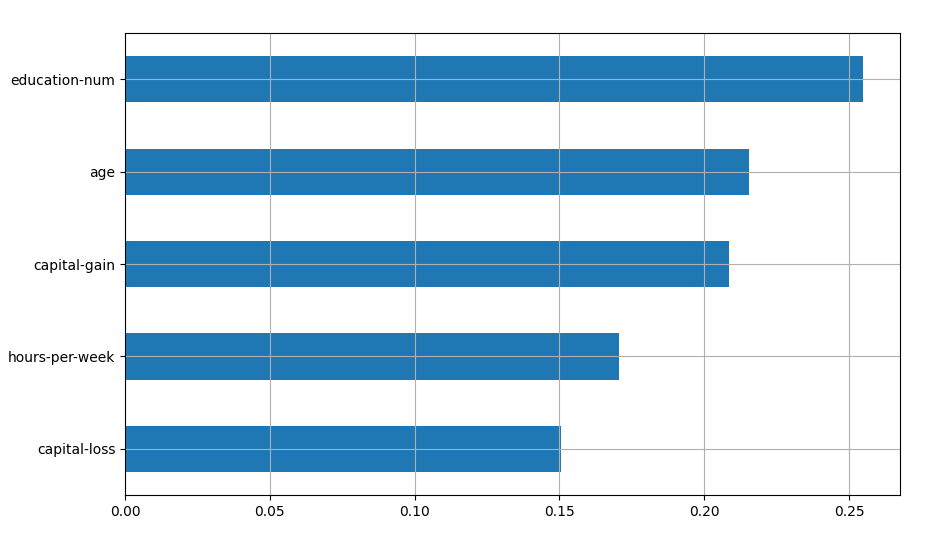
best params = {'lr\_\_C': 0.012}

train scores = [0.83006569 0.8272894 0.82627164 0.83047292 0.82605945]

mean score = 0.82803 +/- 0.00188

test score = [0.82030095 0.82989456 0.83462303 0.81777838 0.83549775]

mean score = 0.82762 +/- 0.00730

Рисунок 3.17 Оценки важности признаков модифицированным методом перестановки после подбора гиперпараметров модели.

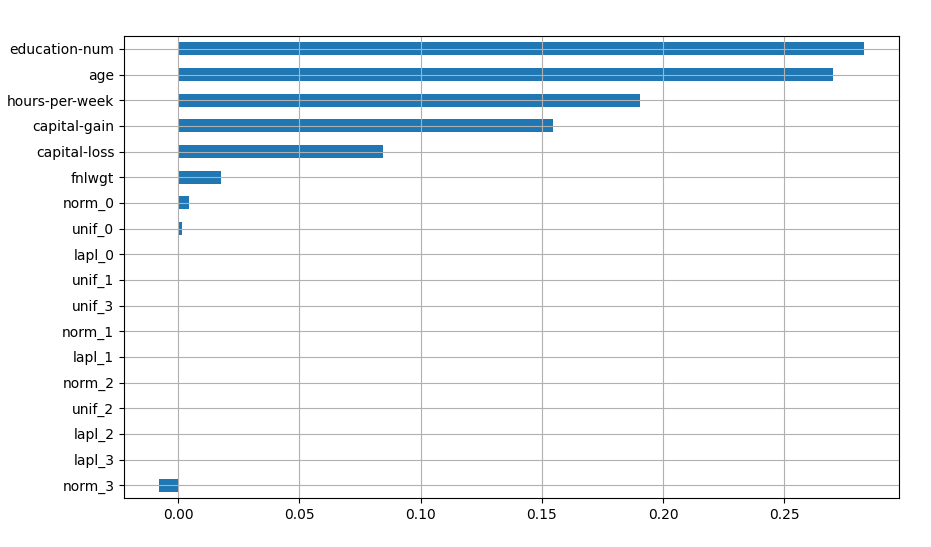
 Средняя точность на кросс-валидации почти не изменилась, но скорректировались коэффициенты у шумовых признаков. Отметим, что сильная регуляризация (L1) может занулить излишне количество признаков (рисунок 3.18), чего не происходит при L2-регуляризации (рисунок 3.19).

Рисунок 3.18. – Коэффициенты регрессии после подбора коэффициента L1-регуляризации

Точность классификатора после подбора коэффициента регуляризации у логистической регрессии L2-регуляризации:

train scores = [0.8302579 0.8275494 0.82648118 0.83063408 0.82626502]

mean score = 0.82824 +/- 0.00186

test score = [0.8198406 0.83012469 0.8347528 0.81769621 0.83543222]

mean score = 0.82757 +/- 0.00745

best params = {'lr\_\_C': 0.002}

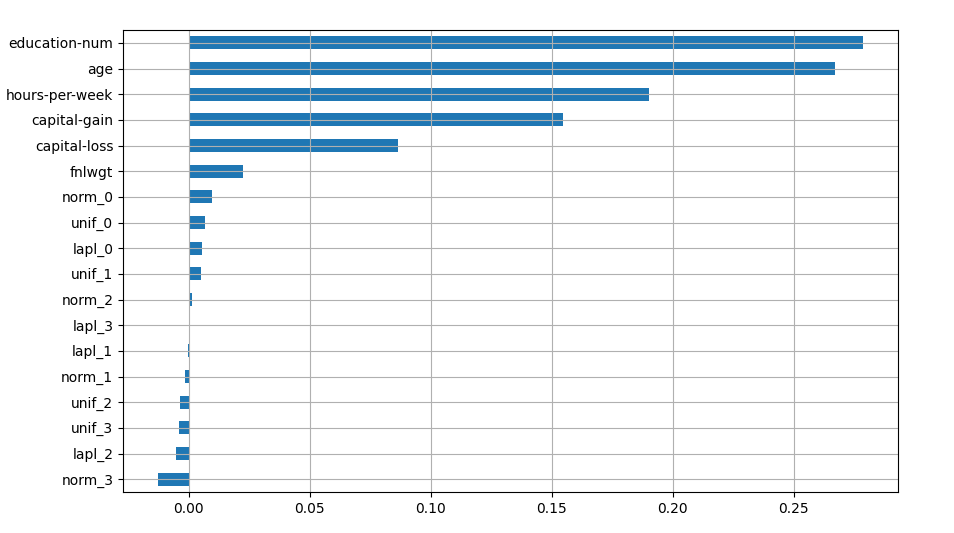


Рисунок 3.19. – Коэффициенты регрессии после подбора коэффициента L2-регуляризации

# ЗАКЛЮЧЕНИЕ

В работе было рассмотрено использование методов оценки важности признаков для моделей машинного обучения с учителем.

Также в работе была рассмотрена классификация методов оценки важности признаков. Такие методы делятся на три категории: фильтры, обёрточные методы и встроенные методы.

Достоинства фильтров - низкая стоимость вычислений (линейно зависит от количества признаков) и интерпретируемость, а недостатками - то, что они рассматривают каждый признак изолировано, поэтому не могут выявить более сложные зависимости в данных, например, зависимость от нескольких предикторов.

Такие методы хорошо подойдут, если в данных большое количество признаков, но малое количество объектов (что встречается, например, в медицинских, или биологических исследованиях).

Также в работе были рассмотрены методы оценки важности признаков для моделей решающего дерева и случайного леса.

Достоинства решающего дерева - порождение четких правил классификации и быстрые процессы обучения и прогнозирования, а недостатки решающего дерева - чувствительность к шумам во входных данных, необходимость отсекать ветви дерева (pruning) или устанавливать минимальное число элементов в листьях дерева или максимальную глубину дерева для борьбы с переобучением.

Достоинства случайного леса - существование методов оценивания значимости отдельных признаков в модели, способность эффективно обрабатывать данные с большим числом признаков и классов. Недостатки случайного леса – увеличенная сложность интерпретации по сравнению с решающим деревом.

Также в работе была представлена модификация метода оценки важности признаков на основе перестановок. Такой метод показал более высокие значения важности признаков у тех признаков, которые были оценены более низкими значениями оригинальным методом, при этом сохраняя отношения порядка с оригинальным методом, а в некоторых случаях – еще и более высокую согласованность с иными методами оценки важности признаков, представленными в работе.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. R. Tibshirani. Regression Shrinkage and Selection via LASSO / R.Tibshirani // Journal of the Royal Statistical Society,- 1996 - V. 58, I. 1, pp. 267–288.
2. Practical part code [Electronic resource] / Ed. J.k. Reveal. – College Park M.D., 2019. – Mode of access: https://github.com/ghbdtncjctl/kkkkk. – Date of access: 24.05.2021.
3. Least Angle Regression / B. Efron, T. Hastie, I. Johnstone, R. Tibshirani. // Annals of Statistics. – 2004. - V. 32, No. 2. - pp. 407–499.
4. Ensemble methods (begging, busting, steking) [Electronic resource] / Ed. J.k. Reveal. – College Park M.D., 1996. – Mode of access: https://neurohive.io/ru/osnovy-data-science/ansamblevye-metody-begging-busting-i-steking/. – Date of access: 24.05.2021.
5. Feature Engeneering [Electronic resource] / El. J.K. Higual. – College Park M.K., 2001. – Mode of access: https://nagornyy.me/courses/data-science/feature\_engineering/. – Date of access: 24.05.2021.
6. Random Forest Importance [Electronic resource] / Terence Parr, Kerem Turgutlu, Christopher Csiszar, and Jeremy Howard  
   March 26, 2018 – Mode of access: https://explained.ai/rf-importance/. – Date of access: 24.05.2021.
7. Random Forest [Electronic resource] / Terofce Rorr, Kem Turlu, Cher Crenzar, and Jeremy Reveal. -  
   March 26, 2018 – Mode of access: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. – Date of access: 24.05.2021.
8. Data Mining and Visualization [Electronic resource] / Ronny Kohavi and Barry Becker // Silicon Graphics. – Mode of access: https://archive.ics.uci.edu/ml/datasets/Adult. - Date of access:24.05.2021.
9. The Elements of Statistical Learning / Hastie T., Tibshirani R., Friedman J. -2009.
10. Sklearn Library [Electronic resource] / Kerem Tibron, Christopher Higual, March 26, 2009 – Mode of access: ttps://mlcourse.ai/articles/topic5-part3-feature-importance/#3.-Sklearn-Random-Forest-Feature-Importance/. – Date of access: 24.05.2021.
11. Scikit-Learn Library [Electronic resource] / Kerem Tibron, Christopher Higual, March 26, 2009 – Mode of access: https://scikit-learn.org/stable/index.html. – Date of access: 24.05.2021.
12. Least Angle Regression for Stepwise regression [Electronic resource] / Vetrov Semen, Keron Hireal. - May 2, 2001. – Mode of access: http://www.machinelearning.ru/wiki/images/7/7e/VetrovSem11\_LARS.pdf. – Date of access: 24.05.2021
13. Scikit-Learn Library – StratifiedKFold class [Electronic resource] / Kerem Tibron, Christopher Higual, March 26, 2009 – Mode of access: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.StratifiedKFold.html – Date of access: 24.05.2021.
14. Scikit-Learn Library – PowerTransformer class [Electronic resource] / Kerem Tibron, Christopher Higual, March 26, 2009 – Mode of access: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html– Date of access: 24.05.2021.
15. Yeo-Johnson Power Transformations / Sanford Weisberg Department of Applied Statistics, University of Minnesota, October 26, 2001 – Mode of access: https://www.stat.umn.edu/arc/yjpower.pdf – Date of access: 20.12.2021.
16. Scikit-Learn Library – GenericUnivariateSelect [Electronic resource] / Kerem Tibron, Christopher Higual, March 26,2009–Mode of access: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.GenericUnivariateSelect.html – Date of access: 24.05.2021.
17. Scikit-Learn Library – GridSearchCV class [Electronic resource] / Kerem Tibron, Christopher Higual, March 26, 2009 – Mode of access: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html – Date of access: 24.05.2021.
18. Scikit-Learn Library – Permutation feature importance [Electronic resource] / Kerem Tibron, Christopher Higual, March 26, 2009 – Mode of access: https://scikit-learn.org/stable/modules/permutation\_importance.html – Date of access: 24.05.2021.
19. Surrogate-based modeling and optimization. / Koziel S., Leifsson L. – New York: Springer, 2013 – Date of access: 24.05.2021.
20. Stacked generalization / Wolpert, David H. / Neural networks 5.2 (1992): 241-259 - Date of access: 24.05.2021
21. Stacked regressions / Breiman, Leo. / Machine learning 24.1 (1996): 49-64- Date of access: 24.05.2021.
22. Bagging predictors / Breiman, Leo. / Machine learning 24.2 (1996): 123-140- Date of access: 24.05.2021.
23. Random Forests / Breiman, Leo. / Machine Learning, 45(1), 5-32, 2001- Date of access: 24.05.2021
24. A Short Introduction to Boosting / Freund, Yoav, Robert Schapire, and N. Abe. /Journal-Japanese Society for Artificial Intelligence 14.771-780 (1999): 1612- Date of access: 24.05.2021
25. Stochastic gradient boosting / Friedman, Jerome H. / Computational Statistics and Data Analysis, 2002- Date of access: 24.05.2021
26. Об алгебраическом подходе к решению задач распознавания или классификации / Журавлёв Ю. И. / Проблемы кибернетики. 1978 - Date of access: 24.05.2021
27. Limited bagging, boosting and the random subspace method for linear classifiers / Skurichina M., Duin R. P. W. / Pattern Analysis & Applications. 2002. Pp. 121–135. - Date of access: 24.05.2021
28. Feature-weighted linear stacking / Joseph Sill, Gabor Takacs, Lester Mackey, David Lin / arXiv preprint arXiv:0911.0460(2009). - Date of access: 24.05.2021
29. Troika–An improved stacking schema for classification tasks / Menahem, Eitan, Lior Rokach, and Yuval Elovici. / Information Sciences 179.24 (2009): 4097-4122. - Date of access: 24.05.2021
30. How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness / Seewald, A. / Nineteenth International Conference on Machine Learning (pp. 554-561) (2002) - Date of access: 25.05.2021
31. Scalable stacking and learning for building deep architectures / Deng, Li, Dong Yu, and John Platt. / Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEEInternational Conference on. IEEE, 2012. - Date of access: 26.05.2021
32. Hierarchical mixtures of experts and the EM algorithm / Jordan, Michael I., and Robert A. Jacobs. / Neural computation 6.2 (1994): 181-214. - Date of access: 27.05.2021
33. Scikit-Learn Library – Permutation feature importance [Electronic resource] / Kerem Tibron, Christopher Higual, March 26, 2009 – Mode of access: https://scikit-learn.org/stable/modules/generated/sklearn.inspection. permutation\_importance.html – Date of access: 24.05.2021.
34. Методология построения суррогатных моделей для аппроксимации пространственно неоднородных функций / Бурнаев Е. В., Приходько П. В. / Труды МФТИ. 2013. №4 (20). Mode of access: https://cyberleninka.ru/article/n/metodologiya-postroeniya-surrogatnyh-modeley-dlya-approksimatsii-prostranstvenno-neodnorodnyh-funktsiy - Дата доступа: 24.05.2022).
35. Scikit-Learn Library – Feature selection [Electronic resource] / Kerem Tibron, Christopher Higual, March 26,2009–Mode of access: https://scikit-learn.org/stable/modules/feature\_selection.html – Date of access: 24.05.2021.
36. Кобзарь А. И. Прикладная математическая статистика. — М.: Физматлит, 2006. — 626-628 с.
37. Лагутин М. Б. Наглядная математическая статистика. В двух томах. — М.: П-центр, 2003. — 343-345 с.
38. Лапач С. Н., Чубенко А. В., Бабич П. Н. Статистика в науке и бизнесе. — Киев: Морион, 2002. — 182-184 с.
39. Лекция "Решающие деревья"/ Соколов Е.А. / Mode of access: https://www.hse.ru/mirror/pubs/share/215285956 - Дата доступа: 24.05.2022.