МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

(национальный исследовательский университет)»

ОТЧЕТ ПО КУРСОВОЙ РАБОТЕ

по дисциплине «Статистические методы обработки данных»

Тема: «Методы кодирования категориальных признаков в задаче бинарной классификации»

Выполнил cтудент группы М80-101М-21

         Фейзуллин К.М.

Руководитель: канд. физ.-мат. наук, доцент кафедры 804 Платонов Е.Н.

Оглавление

[**Введение** 3](#_Toc90216336)

# **Введение**

Целью работы является исследование влияния методов кодирования категориальных признаков в задаче бинарной классификации на функционал качества работы прогнозирующей математической модели.

В эру больших данных и расцвета автоматизированных алгоритмов обработки информации и прогнозирования важную роль занимают сами методы обработки. Помимо целочисленных и вещественных признаков нередко приходиться сталкиваться и с категориальными данными, которые требуют особый подход для дальнейшей работы с ними.

Актуальность данного исследования состоит в том, что до сих пор точно неизвестно, для какого рода задач и для каких сущностей переменных то или иное кодирование покажет наилучший результат.

Причем, данная задача бинарной классификации нередко используется для решения разного типа проблем, будь то прогноз подорожания или роста акции, решение о предоставления кредита или ипотеки, определение вредоносности программного обеспечения по показателям системы. Как можно понять, данная задача имеет множество приложений и в том числе выгодных с экономической точки зрения.

Но очень важно учитывать, что качество работы математической модели имеет зависимость с качеством данных, которые обрабатываются алгоритмом прогноза.

Поэтому нахождение оптимальных методов кодирования напрямую скажется на качестве работы прогнозирующей модели.

# **Постановка задачи**

Целью задачи является исследование методов кодирования категориальных признаков в задаче бинарной классификации. Функционалом качества работы алгоритма бинарной классификации является метрика ROC AUC (Receiver Operating Characteristic Area Under Curve), которая представляет из себя интегральную сумму от пороговой функции отношения доли верно определенных положительных меток и доли ложно определенных положительно меток.

Необходимо используя различные способы кодирования категориальных признаков и через создание новых признаков добиться оптимального результата.

Все варианты кодирования (различные наборы признаков) и соответствующую точность решения задачи будут занесены в сводную таблицу.