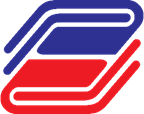
**Министерство науки и высшего образования Российской Федерации**

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УПРАВЛЕНИЯ»**



|  |  |
| --- | --- |
| Институт | Информационных систем |
| Кафедра | Математических методов в экономике и управлении |

**Отчет по проектной работе в рамках дисциплины**

**“Методы и технологии машинного обучения”**

|  |  |
| --- | --- |
| Тема проекта | Прогнозирование устройства на работу соискателя согласно данным портала «Работа в России»: обработанные и объединенные сведения о вакансиях, резюме, откликах и приглашениях портала trudvsem.ru |
|  |  |
| (тема проекта) | |
|  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Направление подготовки | 01.03.02 |  | Прикладная математика и информатика | |
| (код) |  | (наименование) | |
| Образовательная программа | Прикладная математика и информатика | | |
| (название образовательной программы) | | |
| Обучающийся | Челышев Михаил Сергеевич  Шилкин Степан Владимирович | | |
| (Фамилия, Имя, Отчество) | | |
| III курс, 1 группа |  | | |
| (курс, номер группы) |  | | |
| Руководитель | Денисова А.И. | | |
| (ученая степень, звание, Фамилия и Инициалы) | | |
|  |  | | |

Москва, 2023

**СОДЕРЖАНИЕ**

Оглавление

[РАЗДЕЛ 1. ВВЕДЕНИЕ. АКТУАЛЬНОСТЬ ТЕМЫ И ПОСТАНОВКА ПРОБЛЕМЫ 2](#_Toc132772639)

[РАЗДЕЛ 2. ПРАКТИЧЕСКАЯ ЧАСТЬ ИССЛЕДОВАНИЯ. ПРИМЕНЕНИЕ МЕТОДОВ DATA SCIENCE ДЛЯ ДОСТИЖЕНИЯ РЕЗУЛЬТАТОВ 3](#_Toc132772640)

[РАЗДЕЛ 3. ЗАКЛЮЧЕНИЕ 7](#_Toc132772641)

[СПИСОК ИСТОЧНИКОВ 8](#_Toc132772642)

# **РАЗДЕЛ 1. ВВЕДЕНИЕ. АКТУАЛЬНОСТЬ ТЕМЫ И ПОСТАНОВКА ПРОБЛЕМЫ**

В современном обществе одним из важнейших аспектов является трудоустройство, которое напрямую влияет на благополучие граждан и уровень экономического развития страны. В условиях повышенной конкуренции на рынке труда важно эффективно распределить доступные ресурсы, учитывая потребности работодателей и соискателей. В этом контексте актуальность темы исследования "Прогнозирование устройства на работу соискателя согласно данным портала «Работа в России»: обработанные и объединенные сведения о вакансиях, резюме, откликах и приглашениях портала trudvsem.ru" неоспорима.

Портал "Работа в России" (trudvsem.ru) представляет собой федеральный государственный сервис, который предоставляет информацию о вакансиях, резюме, откликах и приглашениях на работу. Важно не только агрегировать эту информацию, но и использовать её для прогнозирования трудоустройства соискателей.

Целью данного исследования является разработка методов и моделей прогнозирования трудоустройства соискателей на основе анализа данных с портала "Работа в России". Для достижения этой цели необходимо решить следующие задачи:

1. собрать и обработать данные с портала "Работа в России" (trudvsem.ru);

2. исследовать взаимосвязи между различными параметрами, влияющими на успешность трудоустройства;

3. разработать и оптимизировать модели машинного обучения для прогнозирования вероятности трудоустройства соискателей на основе анализа доступных данных.

В ходе исследования будут использоваться различные методы анализа данных, машинного обучения и статистической обработки, что позволит сделать предложение относительно прогнозирования трудоустройства соискателей. Результаты данного исследования могут быть полезны для государственных органов, работодателей, а также для соискателей, желающих улучшить свои шансы на успешное трудоустройство.

# **РАЗДЕЛ 2. ПРАКТИЧЕСКАЯ ЧАСТЬ ИССЛЕДОВАНИЯ. ПРИМЕНЕНИЕ МЕТОДОВ DATA SCIENCE ДЛЯ ДОСТИЖЕНИЯ РЕЗУЛЬТАТОВ**

В качестве исходных данных были представлены строки с резюме соискателей на работу и информацией – был ли данный соискатель принят, отклонён или его резюме ещё находится на рассмотрении (рис. 1), на которых мы и применяем методы анализа данных. Основными используемыми инструментами для работы стали модули pandas, scipy, matplotlib, seaborn для разведочного анализа данных (EDA, Exploratory Data Analysis), sklearn для обучения моделей классификации.

В качестве метода исследований мы решили использовать сразу два метода: логистическая регрессия и дерево решений с целью сравнения получившихся результатов и принятие лучшего.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 1 – фрагмент предоставленных данных

Для начала необходимо исключить лишние признаки, провести очистку пропущенных и пустых данных, а также сгенерировать новые при необходимости.

Рассчитав количество пропущенных значений в каждом столбце, мы убираем признаки, в которых это количество крайне велико относительно общего количества строк. Из оставшихся мы экспертно выбираем кажущиеся значимыми признаки, исходя из специфики данных и количества уникальных значений в каждом столбце. Получаем объясняемый признак – inner\_info\_status, объясняющие признаки: birthday, inner\_info\_status\_rate, salary, gender, industry\_code, busy\_type, region\_code.

Так как данные сайта состоят из примерно 99% принятых вакансий, для выявления признаков, являющихся значимыми для объясняемого признака inner\_info\_status, мы сначала исключаем из данных строки таким образом, чтобы в итоге количество принятых вакансий было равно количеству отклонённых.

Затем мы рандомно делим получившуюся таблицу на обучающую и тестовую выборки с распределением данных соответсвенно равным 80% и 20% из исходной таблицы. Исключаем строки, в которых резюме ещё находится на рассмотрении, так как в этом случае нельзя отнести его к одному из двух значений: «0» или «1», необходимые для используемых нами методов.

Значения объясняемого признака inner\_info\_status заменяем на «1» для принятых резюме и «0» для не принятых, а также заменяем все пропущенные значения во взятых нами признаках на «Не указано». Таким образом мы получили таблицу данных, с которой можно работать.

Так как значения признаков в основном строковые, а те из них, которые являются числовыми не представляют из себя ряд, коррелированный с объясняемой переменной, необходимы некоторые манипуляции с данными:

1. для признака salary: находим максимальное и минимальное значения из исходной таблицы, делим получившийся промежуток на тысячу частей и относим каждое из значений salary из тестовой выборки к одному из получившихся делений; создаём специальный словарь, ключи в котором будут названиями делений; если получается так, что в каком то делении нет строк, то присваиваем данному ключу в словаре значение, равное среднему из получившихся значений в данном словаре, в обратном случае присваиваем рассчитанную вероятность быть принятым для данного деления на основе распределённых по делениям строк; также распределяем данные из тестовой выборки по уже заданным делениям и заменяем значения признака в каждой строке на значение словаря в ключе для определённого деления;

2. для признака birthday: присваиваем значению столбца birthday значение «Не указано» для всех строк, в которых в столбце birthday\_mistake находится значение «1», затем создаём специальный словарь, в котором ключами будут уникальные значения признаков и присваиваем им и значениям каждой строки признака рассчитанную вероятность быть принятым для данного уникального признака на основе распределённых по ним значений; также заменяем значения признака из тестовой выборки в каждой строке на значение словаря в ключе для определённого уникального значения признака; если получается так, что в каком то делении нет строк, то присваиваем данному ключу в словаре значение, равное среднему из получившихся значений в данном словаре;

3. для остальных признаков: создаём специальный словарь, в котором ключами будут уникальные значения признаков и присваиваем им и значениям каждой строки признака рассчитанную вероятность быть принятым для данного уникального признака на основе распределённых по ним значений; также заменяем значения признака из тестовой выборки в каждой строке на значение словаря в ключе для определённого уникального значения признака; если получается так, что в каком то делении нет строк, то присваиваем данному ключу в словаре значение, равное среднему из получившихся значений в данном словаре.

Для прогнозирования вероятности устройства на работу следует заменить исследованные показатели на значения ключей из словарей для каждого из признаков по вышеуказанному алгоритму.

Далее мы считаем корреляцию объясняющих переменных с объясняемой переменной (рис. 2).

Изображение выглядит как текст, письмо

Автоматически созданное описание

Рисунок 2 – point-biserial корреляция объясняющих переменных с объясняемой переменной

Так как корреляция для признаков gender и busy\_type крайне мала, мы исключили их из рассматриваемых данных.

По получившимся данным проводим логистическую регрессию и дерево решений (рис. 3, рис. 4), результаты которых и являются решениями поставленной задачи.

Изображение выглядит как текст, чек

Автоматически созданное описание

Рисунок 3 – результат логистической регрессии

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 4 – результат дерева решений

# **РАЗДЕЛ 3. ЗАКЛЮЧЕНИЕ**

Мы исследовали факторы, влияющие на принятие на работу соискателя, построили логическую регрессию и дерево решений на обучающей выборке и проверили результаты на тестовой выборке. Таким образом мы получили инструмент для прогноза устройств на работу соискателя согласно данным портала «Работа в России».

Исходя из результатов оценки обоих методов, итоговым мы делаем дерево решений, так как её результат оказался более точным (точность логистической регрессии – 0,6571621, точность дерева решений – 0,721721).

# **СПИСОК ИСТОЧНИКОВ**

1. Портал "Работа в России": [сайт]. URL: trudvsem.ru
2. Используемые данные: [сайт]. URL data-in.ru/data-catalog/datasets/186/#dataset-overview