**Кейс «Анализ факторов, влияющих на употребление табака в США»**

**Аннотация.** Курение, несмотря на все существующие предупреждения о его вредности, продолжает оставаться одной из самых распространенных вредных привычек среди населения США. Каждый год тысячи людей умирают от заболеваний, связанных с курением. Изучение причин курения и поиски эффективных способов борьбы с этой привычкой являются важными задачами в области общественного здравоохранения. Программы по борьбе с курением, включающие информационные кампании, консультации и другие меры, могут помочь людям бросить курить и избежать многих заболеваний, связанных с курением. Поэтому изучение данных об употреблении табака и поиск способов борьбы с курением являются важными шагами на пути к более здоровому обществу.

**Цель:** проанализировать данные по употреблению табака в США. Выявить основные факторы, влияющие на уровень употребления табака и предложить прогноз (построить регрессионную модель) по уровню употреблению табака в США в зависимости от выявленных факторов.

**Техническое задание:** требуется проанализировать данные по употреблению табака в США и зависимости количества людей, курящих ежедневно, от таких факторов как год, штат, процент бросивших курить, процент никогда не курящих и процент людей, иногда употребляющих табак, а также дать прогноз по употреблению табака в США. Построить качественную модель и объяснить ее выбор. По построенной модели дать интерпретацию (какие факторы влияют на вероятность курения или бросания курения и определить наиболее важные факторы, которые можно использовать для разработки эффективных программ по борьбе с курением) и дать прогноз по будущим тенденциям и изменениям употребления табака в США. Пояснить полученные результаты.

**Порядок выполнения проекта**

**Шаг 1.** Выполнить подготовку данных к анализу.

1.1. Загрузить данные в датафрейм.

1.2. Убедиться, что все количественные столбцы имеют числовой тип. Если это не так, выполнить преобразование типа столбца к числовому.

1.3. Проверить данные на пропуски. Если в столбцах или строках очень много пропусков (порог выбрать самостоятельно), такие признаки и объекты для анализа малоинформативны, их следует удалить. Оставшиеся пропуски необходимо заполнить любым подходящим методом (средним, модой, медианой, линейной интерполяцией).

***Результатом*** ***первого шага*** должен быть готовый к анализу набор данных в виде датафрейма.

**Шаг 2.** Выполнить предварительный анализ данных.

2.1. Вычислить описательные статистики по колонкам (среднее, моду, медиану, стандартное отклонение, квартили).

2.2. Проверить числовые колонки на наличие выбросов, для этого можно использовать диаграмму «ящик с усами» (boxplot). Если выбросов мало, следует их сгладить любым методом (заменой на 1-ый или 99-ый процентиль, среднее, моду, медиану, линейной интерполяцией). Если выбросов много, и они не сконцентрированы в узком диапазоне значений, а рассеяны по широкому диапазону, тогда можно ничего с ними не делать.

2.3. Проверить колонки на нормальность распределения двумя способами:

1) построить гистограмму и сделать предположение о том, являются ли данные нормально распределенными;

2) выполнить статистический тест на нормальность (уровень значимости выбрать самостоятельно) и убедиться, что выдвинутое ранее предположение о нормальности верно (ошибочно).

2.4. Выполнить корреляционный анализ данных с помощью матрицы корреляции. В случае нормальности всех колонок использовать корреляцию Пирсона, в противном случае – ранговые коэффициенты. Матрицу корреляции отобразить с помощью диаграммы «тепловая карта» (heatmap).

2.5. Выбрать целевую переменную – такой признак, который предположительно зависит от других признаков и может быть смоделирован с их помощью.

2.6. Используя результаты корреляционного анализа, отобрать факторные признаки – те, которые предположительно оказывают наибольшее влияние на целевой и при этом как можно меньше зависят друг от друга.

2.7. Проверить целевой признак на упорядоченность, используя статистический тест (уровень значимости выбрать самостоятельно). Если данные упорядочены, следует их перемешать.

***Результатом*** ***второго шага*** должна быть выдвинутая гипотеза о том, что целевая переменная зависит от факторных переменных и может быть смоделирована с их помощью с применением модели линейной регрессии.

**Шаг 3.** Выполнить моделирование.

3.1. Разбить данные на обучающую и тестовую выборки (соотношение выбрать самостоятельно, обучающая выборка должна быть больше).

3.2. На обучающей выборке построить модель множественной линейной регрессии, используя в качестве целевой и факторных переменных отобранные ранее столбцы.

3.3. Вычислить коэффициент детерминации модели на обучающей выборке. Убедиться, что модель значима по критерию Фишера (уровень значимости выбрать самостоятельно). Если коэффициент детерминации модели низок (меньше 0,7) и (или) модель незначима, попробовать другие комбинации факторных признаков из отобранных ранее. Выбрать такой набор признаков, при котором модели будет наиболее высок и модель при этом окажется значимой.

3.4. Дать интерпретацию коэффициентам модели.

3.5. Построить диаграмму рассеяния, отображающую зависимость целевой переменной от любой из факторных. Наложить линейный график регрессии (не забыв упорядочить данные по оси).

3.6. Выполнить прогнозирование целевой переменной на тестовой выборке. Вычислить.

***Результатом третьего шага*** должны быть выводы о том, насколько хорошо линейная регрессия описывает зависимость целевой переменной от факторных (подтвердилась ли выдвинутая на втором шаге гипотеза), какие факторные переменные оказывают наибольшее влияние на целевую, каково объяснение этого, пригодна ли полученная модель для предсказания значения целевого признака новых объектов.

**Шаг 4.** Представить отчет по проекту в виде документа Word. Отчет должен включать подробное описание всех выполненных действий с анализом, предположениями о зависимостях в данных и т.д. Обязательно приведение кода на языке Python и результатов его выполнения. Также к отчету должна быть приложена ссылка на блокнот с расчетами в Google Colab (или блокнот Jupyter Notebook, или файл с исходными кодами – если использована иная среда разработки на Python).

**Шаг 5.** Подготовить презентацию по полученным результатам (в формате PowerPoint). Презентация должна содержать краткое описание хода выполнения проекта (описание данных, их подготовка, предварительный анализ, выдвинутые гипотезы, моделирование) и выводы, сделанные по итогам анализа (подтвердились ли гипотезы?). Количество слайдов – не более 20.

**Файлы:**

Smoking.csv – файл, содержащий данные об уровне употребления табака в США.

**Данные**представляют из себя таблицу с названиями колонок:

*Year* – год.

*State* – название штата.

*Smoke everyday* – процент людей, которые курят каждый день **(целевая переменная)**

*Smoke some day* – процент людей, которые курят иногда.

*Former smoker* – процент людей, бросивших курить.

*Never smoked* – процент людей, которые никогда не курили.