Universidad de Sonora

Maestría en ciencia de datos



**Productividad de trabajadores**

Integrantes:

Ballesteros Valenzuela, Diana Laura.

Granich Armenta, Aaron Francisco.

Reyes Paz, Melissa.

Rodríguez Martínez, Eliud Gilberto.

**Productividad de trabajadores**

1. **Comprensión del negocio**

1.1 Determinación de los objetivos de la línea de investigación

Lograr manejar cierto tipo de información de trabajadores de una empresa para tratar de comprenderlos y predecir su productividad.

1.2 Valoración de la situación actual

La industria de la confección es una parte importante de la producción manufacturera, del empleo y del comercio en muchos países en vías de desarrollo.

Bangladesh es un país en vías de desarrollo que es el segundo país exportador de prendas de vestir más grande del mundo. Aparte de la perspectiva económica, la industria de la confección es una de las industrias más intensivas en mano de obra del mundo.

Un gran número de personas están directamente involucradas en el sector textil, de la confección y del calzado en todo el mundo y, por lo tanto, la producción en la industria de la confección depende en gran medida de la competencia y la productividad de los empleados.

En las últimas décadas, se ha publicado una cantidad considerable de estudios en la disciplina de la industria de la confección que se ocupa de la mejora de la producción, la mejora de la calidad y la minimización del costo de fabricación, y la reducción de la complejidad de la producción.

También hay estudios en esta disciplina que se han ocupado de la aplicación de técnicas de minería de datos y aprendizaje automático en este sector. Intuitivamente, la minería de datos y el aprendizaje automático es un proceso que se utiliza para descubrir el conocimiento valioso oculto mediante el análisis de una gran cantidad de datos. La industria de la confección produce una gran cantidad de datos a partir de la actividad diaria intensiva en mano de obra y del proceso industrial. Estos datos se pueden utilizar utilizando técnicas informáticas avanzadas para extraer los conocimientos ocultos y mejorar su rendimiento de producción, así como para resolver muchos de sus problemas industriales y de gestión.

Por lo general, las personas buscan productos de alta calidad pero a bajo costo y, por lo tanto, fabricar los productos sin ningún defecto se convirtió en un desafío para los fabricantes para satisfacer la demanda de los clientes.

1.3 Determinación del objetivo DM

Una línea de producción de prendas de vestir consta de varios procesos secuenciales, como el diseño, la confirmación de muestras, el abastecimiento y la comercialización, la planificación de la colocación, la planificación de marcadores, el tendido y el corte, la costura, el lavado, el acabado y el embalaje.

Para cumplir con los objetivos de producción, los ingenieros industriales asociados establecieron estratégicamente un valor de productividad específico para cada equipo de trabajo en el proceso de fabricación.

Sin embargo, es un escenario común el hecho de que la productividad real de los equipos de trabajo no siempre cumpla con su productividad objetivo y esto puede ser debido a varios factores, como la cantidad de empleados en un equipo, la cantidad de trabajo en progreso, horas extra, incentivos monetarios, entre otros. Esta brecha entre la productividad objetivo y la real de los equipos de trabajo a menudo dificulta que la producción se complete a su debido tiempo ya que el resto del proceso se lleva a cabo basándose en números que no son alcanzados.

Este problema puede provocar una gran pérdida financiera para la empresa en cuestión. Por lo tanto, al desarrollar un plan de línea de producción, un ingeniero industrial debe establecer la productividad objetivo de manera precisa y basada en datos reales para que se minimice la brecha entre la productividad objetivo y la real. Si se puede predecir la productividad real de los equipos de trabajo y establecer el objetivo de acuerdo con la predicción, podría minimizarse la pérdida financiera, maximizar la ganancia y aumentar la eficiencia de la producción.

1.4 Realización del plan del proyecto

Un paso esencial para resolver un problema real es el preprocesamiento del conjunto de datos con los que se trabajará de forma correcta. Es uno de los principales factores para obtener un buen resultado del modelo. Sin embargo, ya que se habla de un problema real, los datos se pueden encontrar ruidosos, malformados, incompletos, faltantes y desequilibrados.

Antes de ajustar los datos en un modelo, es necesario hacer que los datos sean adecuados para el mismo para que pueda comprender los patrones subyacentes en los datos con precisión. La fase de preprocesamiento de datos puede incluir una serie de tareas según la forma y la estructura de los datos, como la limpieza de datos, la codificación de características, la selección de instancias, la normalización, la transformación, la extracción y selección de características.

En este estudio, se ha estructurado esta fase de acuerdo con la forma del conjunto de datos, que incluye ingeniería de características, codificación de características, escalado de características y partición de datos, todo esto fue necesario para después llevar a cabo la preparación de los datos, modelado, evaluación e implementación o conclusiones.

1. **Comprensión de los datos**

2.1 Recolección de datos iniciales.

El conjunto de datos recopilados se originó en el departamento de Ingeniería Industrial de una unidad de fabricación de prendas de vestir de una empresa ubicada en Bangladesh. El conjunto de datos recopilados contiene los datos de producción del departamento de costura y acabado de un lapso de tres meses que van desde enero hasta marzo de 2015.

2.2 Descripción de los datos.

Se cuenta con un dataset que consta de 15 variables, las cuales tienen 1,197 instancias cada una y hay un total de 506 valores nulos.

Las variables incluídas en el dataset son las siguientes:

* date: Fecha en MM-DD-AAAA
* Quarter: Una parte del mes. Un mes se divide en cuatro trimestres
* department: Departamento asociado
* day: Día de la semana
* team: Número de equipo asociado a la instancia
* targeted\_productivity: Objetivo de productividad fijado por la autoridad para cada equipo para cada día
* smv: Valor estándar del minuto, es el tiempo asignado a una tarea
* wip: Trabajos en curso. Incluye el número de artículos sin terminar para los productos
* over\_time: Representa la cantidad de tiempo extra de cada equipo en minutos
* incentive: Representa el importe del incentivo financiero (en BDT) que permite o motiva una determinada acción
* idle\_time: La cantidad de tiempo en que la producción se interrumpió debido a varias razones
* idle\_men: El número de trabajadores que estuvieron inactivos debido a la interrupción de la producción
* no\_of\_style\_change: Número de cambios en el estilo de un determinado producto
* no\_of\_workers: Número de trabajadores en cada equipo
* Actual\_productivity: Es el porcentaje real de productividad que han aportado los trabajadores. Va de 0 a 1.

2.3 Exploración de los datos.

Los datos con los que se va a trabajar están organizados en un dataset con 1197 entradas numeradas del 0 al 1196 y un total de 15 columnas. Se cuenta con seis columnas con datos tipo float64, cinco con tipo int64 y cuatro con tipo object.

2.4 Verificar la calidad de los datos.

Analizando los datos, se puede observar que la variable “wip” cuenta con 506 valores nulos.

También que la variable “department” tiene dos valores distintos, los cuales son “sewing” y “finishing”. El primero está presente 691 veces y el segundo 257 y 249 veces, esto se debe a que por un error de escritura (un espacio al final de la palabra) se tomó como un valor diferente.

De igual manera se ve que, según la descripción de las variables, “actual\_productivity” va de 0 a 1, pero sí hay valores que se salen del rango establecido, superando el 1.

De forma resumida, las restricciones que se tienen es que hay valores nulos, datos con tipos que no concuerdan y también valores fuera de rango, a los cuales no se puede dar explicación, por lo que, para dar solución a estos problemas, se fusionaron las columnas “finishing” y “finishing “ para obtener solamente dos valores diferentes en “department”.

Los valores mayores a 1 en “actual\_productivity” se manejaron de dos formas distintas: se eliminaron y también se igualaron a 1. Los valores nulos en la variable “wip” fueron eliminados.

De igual forma, ya que los datos están comprendidos solamente en tres meses del año 2015, se optó solamente dejar el día que se consideró suficiente para el análisis.

1. **Preparación de los datos**

La preparación de nuestros datos para entrenamiento constó de lo siguiente:

1. Primeramente fue transformar nuestros datos no numéricos a numéricos con la ayuda de MultiColumn Label Encoder.
2. Aplicamos PCA para observar si nos podemos deshacer de alguna variable que solo estuviese poniendo ruido en el entrenamiento.
3. Estandarizamos nuestros datos.
4. Y decidimos separar nuestros datos 80% entrenamiento, 20% para testeo.
5. **Modelado**

Decidimos elegir los siguientes algoritmos para entrenar:

* Linear Regression
* XGB
* Decision Tree Regressor

1. **Evaluación**

**Como métricas de evaluación utilizamos 3 conceptos:**

* **Error cuadrático medio(MSE)** : mide la proximidad de una línea de regresión a un conjunto de puntos de datos. Es una función de riesgo que corresponde al valor esperado de la pérdida de error al cuadrado.
* **Error absoluto medio(MAE) :** Media de las diferencias entre los valores predichos y los valores de observaciones reales.
* **R2 score**: Nos dice que tan correlacionadas son las variables. Un valor bajo mostraría un bajo nivel de correlación, lo que significa un modelo de regresión que no es válido.

Eliminando valores mayores a 1 en la productividad actual:

| Algoritmo | MSE | MAE | R2 score |
| --- | --- | --- | --- |
| Linear regression | 0.008 | 0.05 | 0.72 |
| XGB | 0.007 | 0.04 | 0.76 |
| Decision Tree Regressor | 0.007 | 0.035 | 0.77 |

Reemplazando valores mayores a 1 en la productividad actual:

| Algoritmo | MSE | MAE | R2 score |
| --- | --- | --- | --- |
| Linear regression | 0.012 | 0.06 | 0.73 |
| XGB | 0.008 | 0.05 | 0.82 |
| Decision Tree Regressor | 0.019 | 0.06 | 0.57 |

1. **Implantacion**

Con los datos nos pudimos dar cuenta de varias cosas.

La variable de actual productivity está bastante correlacionada con la variable de target productivity, lo que nos dice que tener un objetivo ayuda sin duda a que la productividad sea mayor; también, que nos podamos deshacer de varias variables porque no aportan mucha información de los datos, con ayuda de PCA.

En el futuro podríamos investigar algún otro manejo de los datos, así como intentar nuevos algoritmos y comparar resultados.

Repositorio en github:

[predicting-Workers-Productivity/regression-to-calculate-productivity.ipynb at main · LauraDi/predicting-Workers-Productivity (github.com)](https://github.com/LauraDi/predicting-Workers-Productivity/blob/main/regression-to-calculate-productivity.ipynb)